A Temporal Topic Model for Social Trend Prediction

by

Somayyeh Aghababaei

A thesis submitted in fulfillment for the degree of

Doctor of Philosophy

in the

Department of Electrical, Computer and Software Engineering

Faculty of Engineering and Applied Science

University of Ontario Institute of Technology

August 2017

Abstract

Social media provides increasing opportunities for users to voluntarily share their thoughts and concerns in a large volume of data. While user-generated data from each individual may not provide considerable information, when combined, they include hidden variables which can convey significant events. In this thesis, we pursue the question of whether social media context can provide socio-behavior "signals" for scio-economic index prediction. The hypothesis is that crowd publicly available data in social media, in particular Twitter, may include predictive variables which can indicate the changes of socio-economic indexes.

We developed content-based and user-centric prediction models where the objective is to employ Twitter content to predict whether the rates increase or decrease for the prospective time-frame. In order to collect Twitter data, we developed an activity-based sampling approach to collect credible users. The intention is to target users who are historically active rather than those who do not have enough contributions in the past. In fact, the idea is to decrease activity gaps or missing opinions of users by developing a data collection method, in which active users are selected for retrieving historical tweets.

Since our problem has a sequential order, extracting meaningful patterns from historical tweets involves temporal analysis. Prediction models require to address information evolution, in which data are more related when they are close in time rather than further apart. We introduced a four-phase temporal topic detection model to infer predictive hidden topics. The model includes document partitioning, topic inference, topic selection, and document representation phases. In fact, a dynamic vocabulary is built to detect emerging topics. The extracted topics are compared over time to select more diverse and novel topics in each time consideration. The selected topics as

predictive features are then applied in the proposed prediction models. The prediction models were evaluated for crime prediction in Chicago, Houston, San Francisco, and Philadelphia. The conducted experiments revealed the correlation between features extracted from the content and crime rates directions. The findings indicate that, extracted topics from content of active users achieved better performance compared to other features such as auxiliary ones. In addition, the proposed sampling approach decreased missing opinions, therefore, the prediction performance was increased significantly. Overall, the proposed models in Twitter data collection and temporal topic detection have contributions in user-based sampling approaches and sequential topic detection problems, respectively. The research also provides insight into the correlation of social content and crime trends as well as the impact of social data in providing predictive indicators.

Acknowledgements

I would like to express the deepest appreciation to my supervisor Dr. Masoud Makrehchi for his invaluable guidance, support and immense inspiration during my PhD. You always provided me the best directions, advices, and novel ideas. Thanks for giving me the freedom to explore many interesting research problems and thanks for believing in me and eventually helping me to achieve my goals. Without your patience, knowledge, and encouragement my study would not be achievable.

A sincere appreciation is expressed to Dr. Shahryar Rahnamayan, Dr. Ramiro Liscano, and Dr. Faisal Z. Qureshi who provided me insightful feedbacks on my PhD proposal.

Many thanks go to all my friends and colleagues, Seama Koohi, Zeinab Joudaki, Iullia Chepurna, Glenda Benni, Sanaz Khakpour, Mehrin Gilani, present and past who always cheered my up and motivated me through my journey. I would like to thank my friends at Scilab, Mehran Kamkarhaghighi, Tara Ahmadalinezhad, Eren Gultepe and Neil Seward who contributed to the friendly atmosphere of the lab.

I would like to also gratefully thank the Ontario Trillium Scholarship (OTS) and Natural Sciences and Engineering Research Council (NSERC) for their generous financial support in the past four years on my PhD.

Many thanks to my family, my beloved parents, and parents-in-law for their continued and unconditional support and love. My parents are my heroes who made many sacrifices to help me achieve my goals. A very especial thanks go to my sisters and brother in-law, Mahdieh, Mohiedin, Mona, and Sahar who know me in a way that nobody else can . Thanks for your invaluable support and love. No matter how far we are, you are always my soul and inspiration to continue my journey.

And my must heartfelt thanks to a very special person, my husband, Adel A.Zadeh. Without you I could not make this valuable journey happen. I appreciate your understanding and believing in me. You constantly supported me and helped me through

each step of my life. You were the one seeing the best of me and helping me to tackle the challenges.

To my grandmother …

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Owing to the spread of Microblogs, opportunities for individuals to voluntarily share their thoughts, concerns, and opinions have dramatically increased. User-generated content has been widely adopted to both explore the language of crowds and leverage this in real world problem predictions. The vast amount of publicly available content has been utilized to predict real-time notifications [1], social conflicts [2], and public health risks [3]. In fact, content includes those certain patterns that describe diverse communities, opinions, and social behavior. Those certain patterns, captured from the context of all individual users, are the result of using different phrases, expressing various sentiments or having a distinguished language model between users.

In this thesis, we assume that social media context can provide "signals" for predicting socio-economic indexes. While predicting social indexes rely on the availability of historical data, social context provides ideal data resource for many real world predictions. The hypothesis is that publicly available data in social media, in particular Twitter, may provide predictive variables which can indicate future social indexes without being limited to the availability of historical records. Twitter is a better data resource

compared to other Microblogs, since the culture of sharing content with public provides a more relaxed setting. In fact, knowing audiences has significant impact on "self-presentation" for online users[4]. In Twitter, this implies that users share their thoughts with public with less concern for social expectations.

We leverage Twitter data to predict the changes of social indexes. In this thesis, the problem of trend prediction is to detect the directions of the target trends based on the previously posted tweets. In fact, we converted the problem of real world trend prediction to text classification. In contrast to many classification problems, the proposed text classification does not suffer from the lack of annotated data. Training data is generated by annotating voluntarily shared content as learning examples with knowledge inferred from the trend. This model infers labels from the environment, events, metadata or any background knowledge captured from the problem. In fact, the concept of data annotation is similar to other labeling approaches, such as the classic lexicon-based approach [5]. In this approach, polarities or labels are inferred based on a set of vocabularies. Thus inspired, in our prediction models we infer labels based on the changes in the objective trends. The content of collective users is labeled positive or negative if the trend goes up or down in the prospective time-frame, respectively.

However, when dealing with the content of users, activity gaps or the sparsity of users' activities decrease the performance of prediction models. In fact, users who are not constantly active over time raise the issue of missing data, therefore, their content is not credible for Twitter-driven prediction models. Nevertheless, determining the subset of "credible" users is crucial. While the majority of user sampling approaches focus on individuals' static networks, dynamic users' activities are not usually considered, which may result in activity gaps in the collected data. Models based on noisy and missing data can significantly degrade in performance. In this thesis, we investigate how to sample Twitter users in order to produce more credible data for temporal prediction models. We present an activity-based sampling approach where users are selected based on their historical activities on Twitter. The predictability of the collected

content from the activity-based and the random sampling is compared in a user-centric temporal model for trend prediction.

Nevertheless, there are some challenges in exploiting content for trend prediction. Despite having problems with processing the content of tweets, for example abbreviations and the limited number of characters, content has a time varying nature. As an example, in our model, documents come as a stream where terms which appear in one document may no longer be popular in upcoming documents. Therefore, discussion topics have the characteristics of birth and death [6, 7], such that, some words such as "Democrat" and "Republican" are popular in users' discussions before presidential election but it is not certain that they remain as hot topics for a long period of time. In fact, term usage in daily conversation changes based on users' concerns and interests. Therefore, we need to address temporality changes of terms in our prediction model. The idea is to extract emerging topics as the predictive features, rather than static topics where the distribution of terms does not change over time. In this regard, we propose a topic model which builds a dynamic vocabulary to infer emerging topics and fade away vocabularies which are no longer popular.

1.2 Research Questions

In this research, we aim to answer and investigate the following questions:

- Can content – as user-generated data – predict socio-economic trends?
  In some domains, such as crime prediction, the contribution of social data as a freely available data resource is less explored.
- Is the approach of generating training data, with the knowledge inferred from the problem, effective in predicting a trend of interest?
- What are the most informative features for predicting the targeted problem?

- How to tackle the problem of activity gaps on Twitter?
- How to efficiently extract topics for the prediction model over time?

1.3 Research Objectives

The main objectives of this research are to:

- Ascertain the contribution of the content, that is voluntarily shared on Twitter, in real world predictions such as socio-economic indexes.
- Present a user-based sampling approach based on dynamic users' activities to decrease missing opinions for prediction models.
- Develop a temporal topic detection model in which the changes in vocabulary are leveraged to detect emerging topics.

1.4 Research Contributions

The following papers have contributed to the research outcomes:

- Somayyeh Aghababaei and Masoud Makrehchi. "A Temporal Topic-based Approach for Crime Prediction", IEEE Transactions on Cybernetics (To be submitted).
- Somayyeh Aghababaei and Masoud Makrehchi. "A User-based Filtering Approach for Prediction of Crime Trends from Sentiments of Active Users in Twitter", Journal of Social Network Analysis and Mining, Springer (Submitted).

- Somayyeh Aghababaei and Masoud Makrehchi. “Mining Twitter Data for Crime Trend Prediction”, Intelligent Data Analysis. Volume 22(1). January 2018.

- Somayyeh Aghababaei and Masoud Makrehchi. “Activity-based Twitter Sampling for Content-based and User-centric Prediction Models”, Human-centric Computing and Information Sciences, Springer. Volume 7(1). January 2017.

- Somayyeh Aghababaei and Masoud Makrehchi. “Interpolative Self-training Approach for Sentiment Analysis”, the 3rd International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2016).

- Somayyeh Aghababaei and Masoud Makrehchi. “Activity-Based Sampling of Twitter Users for Temporal Prediction Models”, the 3rd International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2016).

- Somayyeh Aghababaei and Masoud Makrehchi. “Mining Social Media for Crime Prediction”, 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI’16).

- Mehran Kamkarhaghighi, Iuliia Chepurna, Somayyeh Aghababaei, Masoud Makrehchi. “Discovering Credible Twitter Users in Stock Market Domain”, 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI’16)

- Somayyeh Aghababaei and Masoud Makrehchi. “Temporal Topic Inference for Trend Prediction”, in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on , vol., no., pp.877-884, 14-17. November 2015.

- Somayyeh Aghababaei and Masoud Makrehchi. “Self-Labeling Approach for Crime Trend Prediction from Twitter data”, Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer International Publishing, 2015. (accepted)

- Iuliia Chepurna, Somayyeh Aghababaei, and Masoud Makrehchi. "How to Predict Social Trends by Mining User Sentiments". Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer International Publishing, 2015. 270-275.

- Somayyeh Aghababaei and Masoud Makrehchi. "Crime Trend Prediction Using Social Data", Social Media and Society International Conference 2015 (SMSociety 15). Toronto.

- Somayyeh Aghababaei and Masoud Makrehchi . "Inferring Influence of Twitter Content on Crime Trend Prediction". Poster at IBM Centers for Advanced Studies: CASCON 2014.

1.5 Structure of the Thesis

This thesis is organized as follows:

Chapter 1 provides the motivation for the research along with research questions, objectives, and contributions.

Chapter 2 discusses a review on conventional crime prediction methods, Twitter driven prediction models, temporal topic detection, as well as Twitter sampling approaches. In each subsection, we also discuss the motivation of this thesis based on the proposed models.

Chapter 3 describes the proposed prediction models as well as the labeling approach. Two models are presented for trend prediction; content-based and user-centric models.

Chapter 4 presents the proposed activity-based sampling approach along with the evaluation of the data collected using the activity-based and the random sampling. The datasets collected from both sampling methods are compared in terms of statistics of the data, activity gaps, and prediction performance.

Chapter 5 presents the proposed temporal topic detection. In addition, the results of prediction performance using temporal topics is discussed. We also show the results of other features such as sentiment, BOW, and auxiliary features.

Chapter 6 provides the summary of the thesis as well as the limitations and challenges of the current study with future works and plans.

Chapter 2

Background and Related works

2.1 Twitter-driven Prediction Models

Twitter with around 140 million of users [8] and 350 millions of tweets per day is considered as a rich source of information to understand and predict users' behavior, public sentiments, and events of interest. There have been enormous efforts in utilizing content captured from Twitter to predict real-time notifications, social conflicts, and public health risks [9–11]. Leveraging user-generated data reveals underlying patterns in different domains. Hale et al. [12] studied the validity of language gap between different locations. In this research, latent factors were extracted from content, generated by individuals from different locations, and they were utilized to detect different communities. In another study, Lotan et al. [13] discussed information flow across different political communities and how they were captured by utilizing Twitter content. In fact, content includes those certain patterns describing diverse communities, opinions, and social behavior. Those certain patterns captured from context of all individuals are result of using different phrases, expressing various sentiments or having distinguished language model between users.

Aforementioned studies utilized Twitter data from three different aspects: users' profiles, users' activities, and content shared by users. Users' profiles, such as the number of followings and followers have been leveraged in many user-centric models dealing with the expertise of users to detect the most influential users in the domain of interest [14–16]. In other studies, the activities of users such as retweets, likes, and other users' footprints were applied in learning models to extract meaningful insights. For instance, using retweets as Word-of-Mouth Marketing [17, 18], tracking the evolution of news sharing [19, 20], or detecting extreme events via retweets' networks [21, 22]. Also leveraging content shared on Twitter attracted many attentions in both user-centric and content-based models. The content refers to tweets which are limited to 140 characters and the context vary from daily activities to important news. The tweets can include URLs referring to other pages or hashtags which reflect the similar topics as other tweets with the same hastags. In user-centric approaches, sentiments of users were inferred from their content to capture significant events such as important diseases [23, 24] and natural disasters detection [25, 26]. In content-based models, opinions of crowd collectively provided predictive signals for prediction models [27, 28]. In this thesis, we also focus on leveraging content shared by users on Twitter to extract hidden variables as predictive signals for targeted problems.

2.2 Temporal Topic Models

Topic models were extensively applied to many text mining tasks either to indicate the similarities between two sets of documents [29, 30] or to visualize high dimensional documents to a set of well-structured variables for exploration [31–33]. Nevertheless, with the increasing number of user-generated data in microblogs, there has been a great demand for topic models for learning meaningful patterns from data. Latent Dirichlet Allocation (LDA) [34] is a widely used probabilistic generative technique in topic modeling that discovers underlying topics from text documents. In LDA, inputs

are a bag-of-words representation of documents and outputs consist of latent topics. A topic in LDA is a multinomial distribution of words in the vocabulary, while a document is a multinomial distribution of topics. In LDA, documents are given as a batch, which builds a static vocabulary for inferring topic distributions.

However, in temporal analysis, where information changes over time and upcoming text documents most likely carry new terms, predefining a fixed vocabulary is not practical and raises many issues. First, topics have proven to have birth and death [6, 35], when extracted from temporal text streams. Therefore, there is a significant need of a dynamic vocabularies over time to address emerging terms in topic inference and fade away vocabularies which are no longer popular. Second, in static LDA, insignificant topics get high chances to be detected. In fact, topics consist of common words more likely generated if the documents are given as a single group for topic inference because of having a broad range of documents (daily aggregation of conversations over a long period of time). Third, topics related to significant social events may not be captured. As an example in one day a high number of tweets can be observed due to a day which coincide with some special events such as presidential election. If documents, collected during a long period of time, are given as a single batch, topics related to a specific event are less likely inferred due to the large number of words observed over time.

A range of different approaches have been investigated in topic modeling for non-temporal problems where topics were identified without consideration of topic shifts [36–39]. In contrast, a number of studies viewed topic identification with time dimension [40–44]. Their approaches in temporal topic detection were related to an online learning process such as Online LDA [45], topics were extracted from the current time slice, and reassigning topics for new documents happened by updating parameters based on previously applied model. Whereas, we aim at detecting temporal topics for training data where topic evolution happens in offline mode. The vocabulary is

assumed to be dynamically updated over time to find new topics and to capture the evolution of previously detected topics.

Although topic modeling has been widely applied in many text stream analysis to detect discussion topics in microblogs [46–48], the contribution of the detected topics were less explored. Many studies evaluated extracted topics manually or they used distance measurements [7, 43, 49] to investigate the effectiveness of their topic models. Other approaches tried to understand the significance of the topics by considering their probabilities [7] or calculating their distances with junk topics [50]. However, the feasibility of the significance of detected topics in different applications was less considered, which is the target of the current study.

2.3 Twitter Sampling

With the increasing number of Twitter users, the volume of tweets have become overwhelming and Twitter sampling, the selection of subset of tweets or users, is particularly relevant. Many sampling techniques were studied ranging from topical [51–53] to user-based approaches [54]. The first set of techniques is topic-based sampling, where specific keywords or hashtags were applied to collect tweets through Twitter API [3, 55]. This group of sampling limits the study around the content of shared topics which are not scalable to many applications. The second group focused on sampling a subset of users from their networks [51, 54, 56]. The drawback behind the latter approach is that, the availability of users' posts over time was not considered. In fact, there is no guarantee that sampled users are active on a daily basis. The historical activities of users are necessary for temporal models where content (in content-based models) [57, 58] or user timelines (in user-centric models) are aggregated considering their timestamps [59–61]. Therefore, the activity gap is the main challenge for temporal prediction models, in which the performance of the models can be degraded.

The most common sampling approach is random sampling using Streaming API, which allows retrieving 1% of real-time data with some specific parameters. There have been many empirical studies dealing with the evaluation of the data sampled from the random sampling with other approaches, including random vs Firehose [62]. This study discussed the situations in which the random sampling had less coverage compared to Firehose. However, when there were more specific parameters such as keywords, random sampling could provide "enough" data as Firehose. In another study [51], the Streaming API was compared with the Expert sampling. The expert users were the users with high number of followers. In this study, content of expert users were compared with random users in terms of trustworthy of their content. It was revealed that expert content contained more diverse and popular topics and included less spam, which has application in many topical extraction models such as breaking news detection. Therefore, we can conclude from previous studies and the recent ones [63] that expert sampling is rich in content and is more valuable for content-based models such as topical models. In fact, Twitter streaming preserves the statistics of the sample size as the whole representative sample, but for content-based models which can benefit from the context, expert sampling is more superior. Hence, Streaming API is highly dependent on the type of coverage and the targeted problem. Many empirical studies evaluated the effectiveness of expert sampling in many dimensions such as trustworthiness, diversity of discussion topics, statistic representative of samples, or sentiment. However, there are many challenges in utilizing content of experts, whose corporate accounts in social media are normally managed by a group of employees, compared to random users. In many applications, ranging from content-based [64] to user-centric [59], opinions of crowds collectively provide predictive signals for prediction models. In fact, by filtering experts we ignore the valuable content coming from crowd and we neglected the vast amount of information contributed by the citizens.

A vast amount of studies prefers network sampling rather than selection of experts based on their popularities. In the network sampling, a subset of users are chosen

from the entire network of collected users for perfect sampling. Different techniques have been applied in recent years, of which Random Walk and Breadth-First Search (BFS)[65] are well-known. However, the major problem with the mentioned techniques is that, these techniques are biased toward high degree nodes similar to expert sampling. A solution to this problem is the traditional Monte Carlo Markov Chain (MCMC), which was proposed by White et al. [54]. They applied a technique based on MCMC and Coupling From The Past (CFTP) to have better convergence in sampling. These methods ignore the activity of users over time, Whereas, in temporal models, the presence of users over time is mostly needed.

In temporal models such as detecting targeted events [23, 66, 67], discovering spatio-temporal topics [68, 69], or tracking users' behaviors over time [70, 71], users' activities or content shared over time are tracked to extract meaningful signals. Therefore, activity gaps or missing opinions can significantly degrade the performance of both content-based and user-centric models. Although many sampling approaches were presented to select a subsets of users and content in static mode, there is a significant need of a sampling approach to address temporal aspect of data. In this thesis, we investigate how to retrieve users to decrease the activity gaps. We also investigate how much retrieved content from sampled users are effective for a temporal prediction model. In fact, we leverage users' profiles to estimate their activities in the past for the selection of the most active users as opposed to experts users.

2.4 Crime Prediction

Crime, which can be defined as any unlawful act punishable by law, not only affects individuals who are involved but the society as a whole. In criminology, crime in all its facets occurs due to a set of situations known as the "social context" [72]. Crime and the risk of being victimized are variants that depend on the social context. Social context in general is viewed by two different dimensions; physical and social. The physical

view refers to the specific geographical locations where crime is more common, such as locations with a higher population and a lower economic status or a limited accessibility to education facilities [73]. However, social dimensions are concerned with socio-psychological factors such as individuals personalities, level of education, students behavior at school, or environment.

Crime analytics in general and crime prediction in specific have drawn the attention of researchers to very diverse fields including law-enforcement and policing, social science, and data mining. The main objective in crime analytics is to help law enforcement agencies to more effectively allocate their scarce resources by predicting criminal movements which requires mining vast amounts of crime data, demographic and socio-economic information, and recently, social data.

Various approaches have been undertaken to deal with crime prediction problems. Conventional techniques, used by law enforcement agencies, were mostly based on historical socio-economic indexes and demographic information. The historical data were collected from the areas of concentrated crime, known as hot-spot maps, and were applied to predict distributions of crime from different natures [74, 75]. The studies are peculiar to a specific location and thus cannot be generalized. To overcome this problem, other techniques were proposed to incorporate background knowledge about spatial features, such as the distance to intersections and highways, schools and businesses, and other information about the neighborhood [76, 77]. Mohler *et al.* [78] proposed a framework that models future crimes as the consecutive to currently committed. However, there was some debate that whether these maps indicate the concentration of all crime types [74]. As an example, taxicab robberies take place in different locations which are not always representative of high crime regions [74]. Another major issue is the lack of data for prediction models. In fact, in conventional methods, historical criminal records must be available for prediction models. Overall, the main drawback of these methods is that they reduce the social context to historical crime records while ignoring socio-behavioral data of the community including

both victims and criminals. In fact, the contextual data can be leveraged as a signal to predict upcoming incidents. Another line of the research considers social fabric of the neighborhood as a key factor, which has an influence on criminal activities [79, 80]. The most recent works studied the predictive power of mobile network activities for the similar problem [81, 82]. All historical data, which is leveraged in the mentioned models, is grouped into three main categories: the location of criminals, the locations they live and appear such as escape route and tourist regions, time and weather, and the criminals' networks.

As discussed earlier, the main characteristic of the crime prediction methods is that they leverage historical data of high crime areas for the prediction model. Despite the fact crimes mostly happen in high crime neighborhoods, the information in a hot-spot map of a small geographical size does not necessarily represent crime rates in bigger communities. Studies showed that in community level, activities in hot-spot maps were not always representative of future crime rates [83–85]. As a result, crime prediction of a specific location can not be easily generalized to the other locations. The location-specific characteristic of hot-spot map models implies that we need to collect enough data from the location of the interest. An alternative approach is to build a generalizing model from a type of data which is freely and publicly available and not restricted to a geographic neighborhood. The social media data has these characteristics in addition to its contextual features that can implicitly carry a socio-behavioral state of the public.

There have been enormous efforts in utilizing micro-blog data to predict real-time notifications, social conflicts, and public health risks [1–3]. In fact, leveraging user-generated data reveals underlying patterns in different domains. Chen *et al.* [86] applied textual content of Twitter in the form of user language to detect name-calling harassment. Joan *et al.* [87] also successfully implemented trend prediction on Twitter. In their work, individual behaviors that were extracted from content of daily tweets were utilized to predict socio-economic indexes. In another study, Hale *et al.* [12]

studied the validity of the language gap between different locations. In this research, the latent factors, extracted from user-generated content, were utilized to detect the communities. Considering other social topics, far too little attention has been paid to the effect of on-line user generated data and its associations with crime prediction. Some studies leveraged density of the data captured from social media in crime prediction. Bogomolov *et al.* [81] explored the predictability of the data coming from mobile phones as what they call "human behavioral data" for crime prediction. Similarly, in another study [88], the frequency of violent mobile messages were compared to the residential population for capturing crime hot-spot. However, in both studies demographic information were exploited while contextual social data were not included in the prediction model.

The idea of applying social data for crime prediction can be observed in the works conducted by Wang *et al.* [36], Gerber [8], and Chen *et al.* [89]. The former is the first one to bring social media context into the problem of crime prediction. Wang *et al.* extracted event-based topics from posted tweets to predict hit-and-run incidents in Charlottesville, Virginia. Although the approach is novel, the source of data is limited to a set of manually selected news agencies and neglected the vast amount of information contributed by the citizens. Also the assumption that content of these posts reflect the most recent local events is not always valid. Finally, it is not clear whether the same predictability will be observed when forecasting incidents, other than hit-and-run. Gerber [8] recently utilized social media data to enhance Kernel Density Estimation (KDE) for crime prediction. Unlike previous authors, Gerber did not impose any restrictions on the source of tweets. He also assessed how much improvement can be achieved by adding topics extracted from Twitter for different crime categories. Similarly, Chen *et al.* [89] utilized the sentiment of Twitter data along with weather condition in KDE for predicting the locations and time of theft crime. This study is limited to the spatial information such as weather data for specific time and regions. In the mentioned studies, KDE as a location dependent technique cannot be

easily generalized to the other cities. There is also some types of crime which are not occurring in vicinity of previous incidents and the population of one area may changed frequently [88].

While most of the research on crime prediction is limited to the specific locations, crime types, communities and users, or focused on the specific events, our proposed approach is one of the first crime prediction approaches that can be generalized to any location. Furthermore, the proposed model learns the directions of changes rather than the occurrence of crimes. The importance of detecting crime trend direction is that policy makers and law enforcement agencies are mostly interested to see if the crime in a neighborhood is declining or not. The other advantage of our approach compared to some of the previous researches is that it works for a wide target range of crime types. Our method does not target any specific communities, keywords, terms, hashtags, and events.

Chapter 3

Prediction Models

3.1 Trend Prediction

Time series analysis deals with developing different methods for three types of movements: Trend, seasonal, and irregular movements. In trend analysis, a long term movements is targeted, which is called trend curve. However, in seasonal variations or cyclic movements, the target is to measure identical behavior of time series during specific period of time such as a month, or a week. In irregular movements, there is no specific pattern repeated over time and random movements will occur. As an example, in event detection models [90, 91], the goal is to detect irregular behavior of time series to drive meaningful reasons of the random changes. In fact, the intention is to find causes of changes. However, the goal of this thesis is finding predictive signals rather than causes of changes. In fact, we focus on trend prediction of time series to predict a long time movements using some predictive variables which are not certainly the causes of changes. They are some indicators which are driven from social media to capture the changes of time series.

Socio-economics index prediction, similar to any other signal prediction, is a challenging task. For example, predicting that 25 incidents of homicide will occur next

day seems impossible. On the other hand, the question " what direction does the trend may take tomorrow" may lead us to some extent to a plausible answer. What we mean from "direction" is the sign of the change at $t(i)$ compared to some reference such as $t(i - \Delta r)$, in which Δr is the lead or lag between the documents and time series. A positive change means, signal has a rising trend while a negative change has the opposite meaning, and obviously zero indicates no change in signal between $t(i - \Delta r)$ and $t(i)$. Figure 3.1 shows an example of time series data with the directions of data intervals.

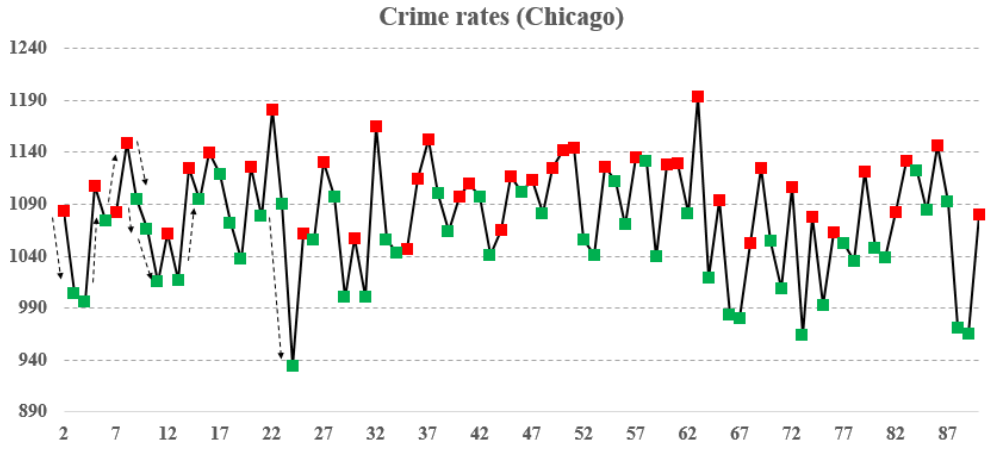


FIGURE 3.1: An example of crime time series with data intervals of high (red) and low (green) crime trend.

We employed crime trend prediction as a targeted problem. The problem of crime trend prediction is considered as a binary classification problem where the objective is to detect the directions of crime trends. Previous studies [8] shown that the classification approach is effective in predicting the occurrence of different crimes, in which class "one" and "zero" are defined as if a specific crime type will occur or not, respectively. However, in our model, classes are directions of crime indexes. In fact, the prediction problem is transformed into a supervised classification task that predicts whether crime rates increase or decrease for the prospective timeframe.

Prior to classification, a set of N training documents of the form $\{(x_1, l_1), (x_2, l_2), ..., (x_N, l_N)\}$ are generated in which x_i is the feature vector of

the i-th document and l_i is its assigned label. For the purpose of creating documents $X = \bigcup_{i=1}^{N} x_i$, two different approaches are applied; concatenation of the content for the content-based approach and the aggregation of user opinions for the user-centric model. In fact, historical tweets are employed into two different models; content-based and user-centric models – they both aim to discover conclusions from user-generated content. In the user-centric model, the content of selected users are aggregated based on users' timelines, to extract meaningful patterns [59], while in the content-based approach, the content of all individuals are combined together with respect to the event of interest [57]. The generated documents are then associated with a set of labels. The labels are inferred from the knowledge obtained from the targeted problem (here crime index), which is the directions of rates in the prospective timeframe. Although the labels are inferred in the same manner for both models (content-based and user-centric), they have different strategies of generating documents. Figure 3.2 shows the framework of the data generation for both models as well as the timeline. After Twitter users were sampled (the sampling approach will be discussed in the next chapter), they are fed to the REST API to retrieve historical timelines of the selected users. The collected data along with the crime rate directions are employed in the content-based and user-centric prediction models, which are discussed in the following subsections.

3.2 Content-based Model

As discussed earlier, in the content-based model, documents are generated based on timestamps of tweets posted by all users without filtering any users. In fact, this model captures collective patterns from the crowd rather than a selected group of users. All observed users are considered as crowd, as opposed to the user-centric approach. In order to generate training examples, a set of temporal document are generated. Let $d_i = \{p_1^{\langle i \rangle}, p_2^{\langle i \rangle}, ..., p_m^{\langle i \rangle}\}$ denotes a document, which consist of a set of posts shared at

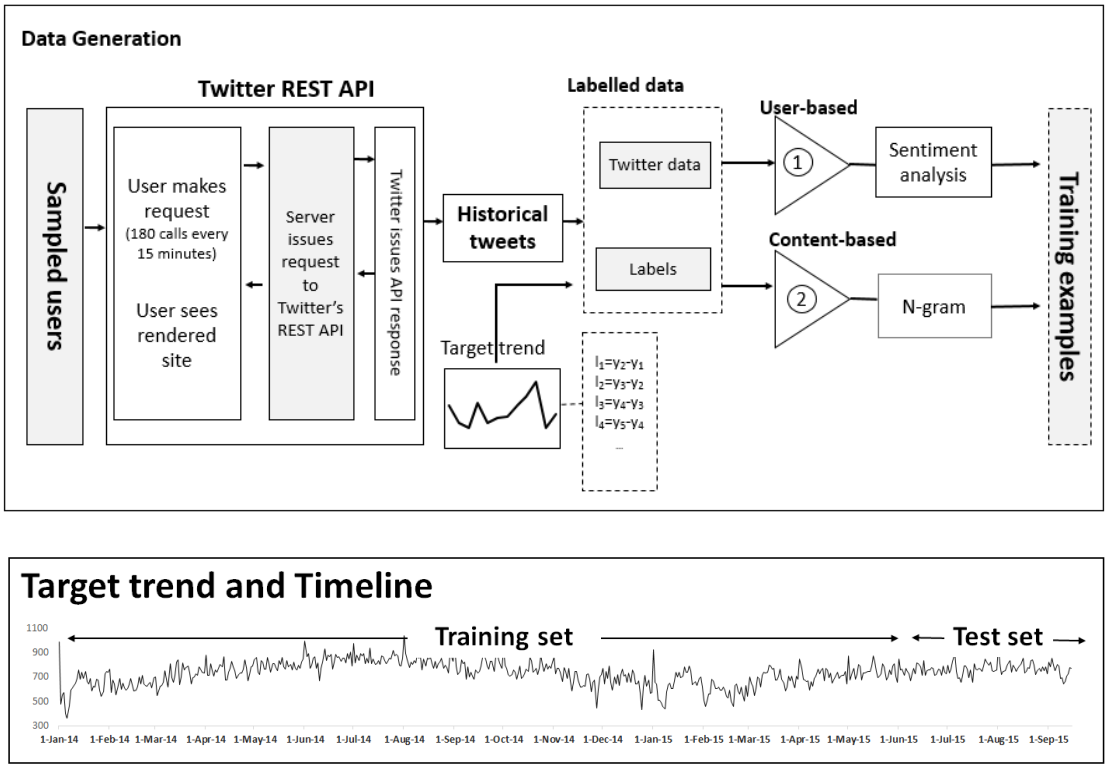


FIGURE 3.2: The framework of the data generation model.

time $t(i)$. Let $D = \{d_1, d_2, ..., d_n\}$ be a set of temporal documents or in general temporal data, which is defined as a state in time. The state is represented by vector of features $d_i = (f_1, f_2, ..., f_{|V|})$, where V is the global vocabulary. Since each state d_i is sampled at time $t(i)$, then $D = \bigcup_{i=1}^{n} d_i$ is the result of n consecutive sampling. One important pre-processing task in time-series data, is smoothing to increase the predictability and to reduce the noise and outliers. Hypothetically, temporal data which is a high-dimensional time-series data can be also smoothed. In our model, each state is represented by a document and a naive smoothing is a rolling averaging algorithm over the temporal documents;

$$x_i^{\langle c \rangle} = \frac{1}{q} \sum_{j=1}^{q} d_{j-q+1}^{\langle c \rangle}, X^{\langle c \rangle} = \bigcup_{i=1}^{n} x_i^{\langle c \rangle}, q = [1, n] \tag{3.1}$$

$$X^{\langle c \rangle} = \bigcup_{i=1}^{n} x_i^{\langle c \rangle} = \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,|V|} \\ f_{2,1} & f_{2,2} & \cdots & f_{2,|V|} \\ \vdots & \vdots & \ddots & \vdots \\ f_{N,1} & f_{N,2} & \cdots & f_{N,|V|} \end{pmatrix} \tag{3.2}$$

where q is the size of aggregation window, d_i is a series of posts shared at time $t(i)$ or in our case the day i, and x_i is a document. All relevant tweets are aggregated into a signal document without targeted filtering. As a result X is an $N \times |V|$ document-term matrix (Eq.3.2) where V is the global vocabulary. The vocabulary V is simply a set of all distinct words appeared in all collected, relevant tweets. Although no keyword search is conducted, a blind filtering including stopword reduction and low-frequent term reduction is applied to the vocabulary. Therefore, x_i is defined as the average of a set of documents from j to day $j - q + 1$, retrospectively.

Several preprocessing tasks such as low frequent term and stopword removal may be applied to x_i. In the content-based approach, documents are represented with terms as features, which are referred to N-gram model without filtering any specific keywords. One might speculate that we must collect keywords to emphasize on offensive language implying a rough context. Nevertheless, content is a rich data which contains valuable hidden variables including activities, topic of discussions, public interests, and sentiments, which might not be necessarily carried by offensive language.

3.3 User-centric Model

In the second model, instead of data aggregation across all users, documents are generated from the individual opinions in different time slots. If a user u_1 has a post at time t and user u_2 also posted something at the same time, the content of each is employed as a unique feature or an user-dependent feature rather than combining them together. Let $Timeline = \{\left(p_1^{\langle u \rangle}, t_1^{\langle u \rangle}\right), \left(p_2^{\langle u \rangle}, t_2^{\langle u \rangle}\right), ..., \left(p_J^{\langle u \rangle}, t_J^{\langle u \rangle}\right)\}$ denotes a timelines of a

user u, where tuple $\left(p_j^{\langle u \rangle}, t_j^{\langle u \rangle}\right)$ represents user u's post j along with its timestamps: $t_1^{\langle u \rangle} < t_2^{\langle u \rangle} < ... < t_J^{\langle u \rangle}$. Post $p_j^{\langle u \rangle} = \{w_1^{\langle u \rangle}, w_2^{\langle u \rangle}, ..., w_k^{\langle u \rangle}\}$, is comprised of tokens $w_k^{\langle u \rangle}$. $V = \bigcup_{u,k} w_k^{\langle u \rangle}$ is a global vocabulary. In order to aggregate tweets based on user timelines, we assume an aggregation window in which user timelines are concatenated as follows:

$$\begin{aligned} d_m^{\langle u \rangle} &= \frac{1}{q} \sum_{j=1}^{q} p_{j-q+1}^{\langle u \rangle}, \quad q = [1, n] \\ x_i^{\langle u \rangle} &= (d_{i,1}^{\langle u \rangle}, d_{i,2}^{\langle u \rangle}, ..., d_{i,M}^{\langle u \rangle}) \\ X\langle u \rangle &= \bigcup_{i=1}^{n} x_i^{\langle u \rangle} \end{aligned} \tag{3.3}$$

where q is the size of aggregation window, M is the total number of users, d_m is a timelines of a user after aggregation, and x_i is a document consist of a series of user timelines. Therefore, features vectors are represented as follows:

$$X^{\langle u \rangle} = \bigcup_{i=1}^{n} x_i^{\langle u \rangle} = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,M} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{N,1} & u_{N,2} & \cdots & s_{N,M} \end{pmatrix} \tag{3.4}$$

where $s_{i,m}$ is the sentiment of the user m, which belongs to document i. Since the idea of this model is considering a sample of users representative of the sentiments of all users, we used Language Inquiry Word Count (LIWC) [92] to detect the sentiment. We extract the positiveness and negativeness scores. Each user is defined by the normalized sentiment scores.

3.4 Generating Labels

In order to infer the labels, let $Y = \{y_1, y_2, ..., y_n\}$ be the target time series whose future values are to be predicted. The time series Y is sampled in time steps $t(i)$, $1 \leq i \leq n$. To convert regression-based prediction into classification, the continuous signal Y has to be mapped into a categorical set which is called the set of labels. There are several techniques to infer labels from a continuous variable such as quantization or direction of changes in rates. Due to the nature of the research, we adopt trend analysis of the continuous rates for labeling:

$$l_i = sgn(y_{i+\Delta r} - y_i), \;\; if \begin{cases} \Delta r > 0: \; lag \\ \Delta r \leq 0: \; lead \end{cases}, L = \bigcup_{i=1}^{n} l_i \tag{3.5}$$

where Δr is the lead or lag from current state (x_i) and target label, l_i is the label at $t = i$ and L is the sequence of labels in n consecutive time steps. After inferring labels, a set of annotated examples is generated by associating high dimensional temporal data to one dimensional target labels inferred from time series of interest,

$$\forall x_i \in X, x_i \rightarrow l_i, D = \{(x_1, l_1), (x_2, l_2) ..., (x_{N-\Delta r}, \; l_{N-\Delta r})\}. \tag{3.6}$$

The objective of the proposed method is to predict whether the trend of interest increases or decreases for the perspective time-frame. Therefore, a set of training data (D) is given to a binary classifier as follows:

$$D = \{(x_i, l_i) | x_i \in R, l_i \in \{-1, 1\}\}, 1 \leq i \leq N - \Delta r \tag{3.7}$$

where in our target problem (crime trend prediction) x_i is learning documents and the label (l_i) is derived from the changes in crime indexes when comparing the current index (i) with the index of ($i + \Delta r$). The inferred label is defined as follows:

$$l_i = \begin{cases} 1 & if rate(i) < rate(i + \Delta r) \\ -1 & \text{otherwise} \end{cases} \tag{3.8}$$

where rate (i) and rate ($i + \Delta r$) are crime index at i and $i + \Delta r$ according to our historical data.

Chapter 4

Activity-based Sampling of Twitter Users

4.1 Introduction

Twitter's public and open nature provides great opportunities for its users to actively participate in sharing their opinions and produce high quality content that is reflective of their tendencies and preferences in their day-to-day life [4]. This vast amount of publicly available user-generated content is applied to many applications ranging from tracking human social behavior [2, 87, 93], detecting events of interest [1, 3, 12], to smart business [94] where domain knowledge is collected through social media. These studies are either concerned with pulling Twitter and aggregating tweets as bulk or tracking historical tweets over time in order to find meaningful patterns for targeted events. The main challenge of the former studies is the limitation of the Twitter API in accessing only 1% of all existing tweets. However, despite this limitation, the latter studies are concerned with retrieving historical timelines of users.

To tackle the above issues of retrieving more tweets beyond the 1% threshold and obtaining historical timelines, topic-based sampling and the REST API are both shown

to be more effective [95, 96]. In topic-based sampling [51], a set of specific keywords or hashtags are applied to collect tweets through the search API. A very substantial problem with this group of sampling is that it is limited to the studies around the content of shared topics, which is not scalable to many applications. In contrast to topic-based models, the REST API can be a user-based scenario, which provides access to user history.

In the case of the REST API, a set of Twitter users are needed in order to retrieve historical tweets. However, the issue of selecting a credible subset of users still remains. The credible users refer to whom they have less activity gaps. Nevertheless, many network-based sampling approaches were studied, which focus on sampling a subset of users from their networks [54] or sampling users based on their popularities [97]. The drawback behind the network-based sampling is that, a set of users are sampled from a static network while ignoring the availability of their posts over time. In fact, there is no guarantee that sampled users are active on a daily basis, which is necessary for temporal models.

In this thesis, we sample Twitter wherein, we propose an activity-based sampling method to retrieve a selection of users for the REST API. In the activity-based sampling, we leverage users' profiles to extract their historical activities. The most active users are assumed as "credible" users for employing in a temporal prediction model. We address two main characteristics in our sampling model: (a) obtaining the most active users, (b) avoiding missing content or activity gaps over time. The term active users does not refer to celebrities, news agencies, or major companies whose corporate accounts in social media are normally managed by a group of employees.

We gathered two samples of Twitter users using our proposed sampling approach and random users. The random users refer to users who post in real-time, which are collected using streaming API. Since Streaming API is widely used approach in many topical and user-based models [98–100], it is important to assess the effectiveness of

the activity-based sampling proposed in this study compared with random sampling using the Streaming API. The selected users from both approaches are employed in the REST API to collect their historical tweets. We compare the content of users, selected from both sampling approaches in different aspects, including statistical properties and predictability in temporal models.

We employ the collected historical content in two temporal prediction models; user-centric and content-based which were discussed in the previous chapter. Both of the aforementioned approaches are considered to be temporal models, which suffer from the challenge of retrieving tweets over time. In a temporal model where content is tracked to detect a set of patterns, the availability of tweets over time significantly affects the model performance. Therefore, temporal models suffer from activity gaps or missing data. We can evaluate the effectiveness of our proposed sampling compared with the random approach in providing more credible content while mitigating the effect of missing content. Overall, the data gathered from the activity-based and random sampling are compared in three main aspects:

(a) **Timelines:** Do the samples provide enough data for the given period of time?

(b) **User activity:** How is the data coverage during the period of interest. Do we observe missing posts over time?

(c) **Content credibility:** How effective is retrieved content for the temporal user-centric and content-based models?

4.2 Sampling Approaches

The objective of this chapter is to present a sampling approach to collect the best representative users for the REST API. In contrast to often used Streaming API, the REST

API can be a user-based approach with less limitation to access Twitter data. Given a set of users, the REST API provides access to historical timelines, with the limitation of at most 3,200 recent tweets for a single user. The main challenge is how to sample Twitter users to avoid the absence of data in historical tweets. Nevertheless, absent data could be inevitable, users do not necessarily share posts on a daily basis. However, as far as possible, to avoid missing opinions in historical tweets, we address some characteristics for the selection of users. In this method, the interest is to find a set of the most active users while showing no bias toward individuals with a high or low number of tweets. We collect users selected by two different sampling strategies; a random approach using the Streaming API and an activity-based sampling which is based on the historical activity of a user. The use of the network-based sampling is not considered in this study due to the nature of the targeted problem. In this study, we are looking for independent opinions, while the network sampling (users and their networks) is biased toward the same opinions.

4.2.1 Random Sampling

As discussed earlier, random sampling is the most common approach to access data streams. In order to obtain random users, we gathered 1% of tweets using Streaming API. The historical timeline of the randomly selected Twitter users are retrieved using the REST API.

4.2.2 Activity-based Sampling

In this method, the interest is to find a set of active users while being unbiased to individuals with very high or low numbers of followers. In our sampling approach, two factors are considered: the period of time a user is active and its daily number of tweets. Since these specifications are not available, we retrieve them from user profiles.

For each tweet, user profile of its author is retrieved, which includes some fields such as: $status\_count$ and $created\_at$. For each user, two main specifications are calculated as follows:

1) The number of days a users is active ($days$). In order to understand for how many days a user is active, we calculate the number of days the user's profile was generated till the current time. A longer period of activity is a primary criteria for the selection. As we track the content of users over time, users who recently became members are ignored.

2) The average number of tweets per day ($tweets\_day$): As this parameter is irretrievable, we leverage the total number of tweets for the user and the number of days a user is active

$$tweets\_day = total\_tweets/days \tag{4.1}$$

where we assume a user has uniform activity behavior. A user is considered active if it has a high number of active days ($days$) and a high number of tweets per day ($tweets\_day$). The active users are classified by using the number of followers to filter out accounts belonging to celebrities, news agencies or major companies. In addition to not being rely on network properties, the proposed sampling method is not topical model for crime prediction. In topical sampling methods are keyword and domain specific, however, we focus on historical number of days and tweets of a specific user.

4.3 Experimental Results

In this section, we evaluate how much the proposed sampling approaches can minimize the lack of data and deliver more informative content. The historical timelines of the selected users from two different approaches; the activity-based and the random sampling are retrieved using the REST API. We evaluate the feasibility of our sampling

approach compared with the random sampling in retrieving historical tweets. We begin with comparing statistical characteristics of data collected from both approaches. The intention is to understand how well data are distributed over time for both sampling approaches. We then evaluate the credibility of the content in the proposed temporal prediction models (discussed in Chapter 3).

4.3.1 Datasets

We tackle crime prediction as a case study. The idea is how to predict crime rate changes from the tweets posted earlier. We collected Twitter data and crime rates from Chicago, Illinois between January 2014 and October 2015. Chicago has been targeted due to its importance as the third populous city in U.S as well as being among top three cities, which attracted the highest number of visitors during 2012.[1] It has been also ranked as the first in the number of murders, second in robbery, and third in the number of property crimes based on an FBI report during 2013.[2]

4.3.1.1 Crime Data

The criminal records were extracted from Chicago Data Portal [3]. This Data Portal is a rich resource providing all reported incidents on a daily basis, which are retrieved from Chicago Police Department system. Information of frequent crimes that have been reported between January 2014 and October 2015 were collected. Each record contains its timestamps, exact location, and crime type. The dates refer to the time of primary investigation, and crime type derived based on the FBI classification system. Figure 4.1 presents the crime rate time series (aggregated rates of all different crime types). The significant changes in the number of incidents are coincided with important holidays

[1] http://en.wikipedia.org/wiki/Chicago
[2] S. Department of Justice, FBI: http://www.fbi.gov
[3] City of Chicago Data Portal: https://data.cityofchicago.org

such as New Year's day and Christmas. However, they might be the result of missing data.

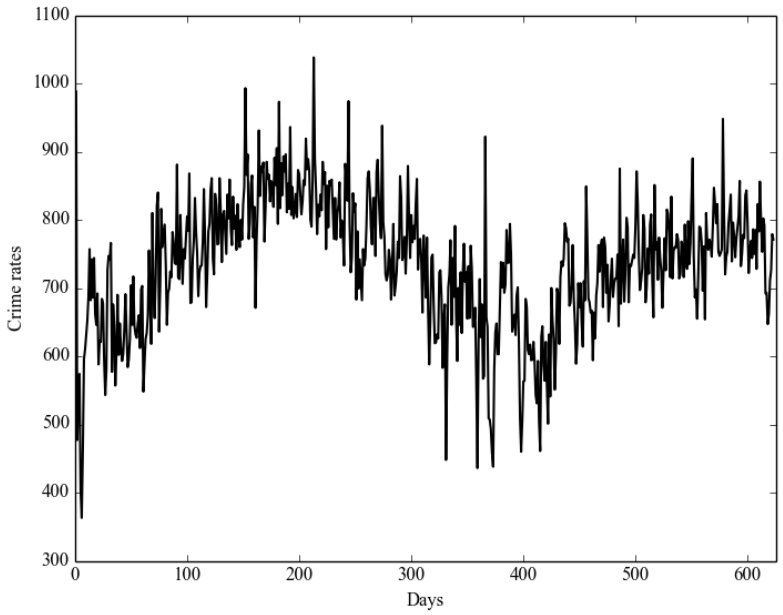


FIGURE 4.1: Daily number of crime rates over 600 days for accumulation of all different crime types.

4.3.1.2 Twitter Data

In order to retrieve historical Twitter data, two sets of Twitter users were collected using the random and activity-based sampling as discussed in previous sections. Historical timelines of the selected users were retrieved and restricted to the same timeframe – between January 1, 2014 and October 1, 2015.

Figure 4.2 presents the number of selected active users over 30 days for the activity-based sampling. The figure indicates two different trends: "Unseen" stands for the number of active users who are selected each day, and " Seen" represents the number of users labeled as active but already selected for that day. As can be observed, the number of new active users who are not detected decreases over time. Due to the increase of repeated users, the process of collecting active users was terminated after almost one month. We applied the REST API to retrieve their historical timelines of the

selected users. Historical timelines of the users were restricted to the same timeframe of crime rates - between January 2014 and October 2015.

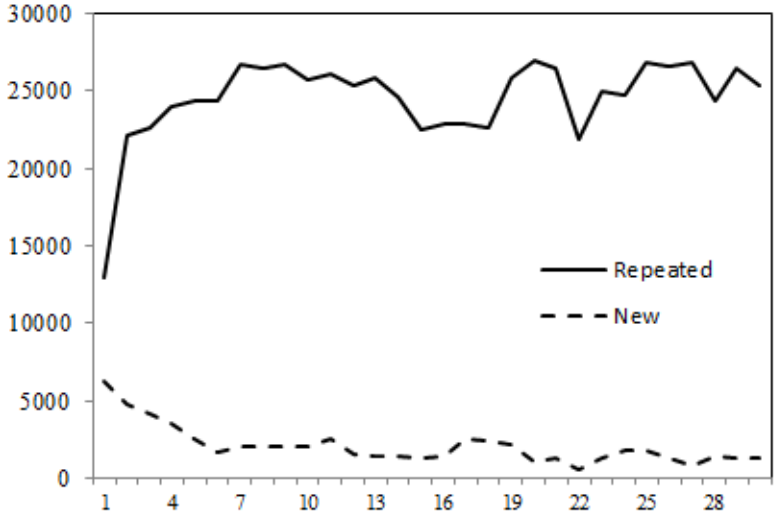


FIGURE 4.2: Daily number of users (y-axis) over 30 days (x-axis).

4.3.2 Comparing Timelines

We compare the number of posts (see Figure 4.3a) and users (see Figure 4.3b) observed on a daily basis from both datasets. The historical tweets obtained from the active and random users are mapped between our consideration period of time using their timestamps, we did not go back more than 600 days because of the low number of activities. As a result, we reached tweets during January 2014- October 2015. Figure 4.3a presents that the daily number of tweets from the active users are higher than tweets of the random users. This can be an asset for content-based models where the availability of content is crucial for the performance of a temporal topical model. For topical models where a set of parameters, such as keywords or hashtags, is retrieved from the collected content, a sampling approach with more coverage might be able to extract more data over time. Figure 4.3b shows the daily number of unique users, defined as those who post at least once per day. From the figure we can observe that the daily number of active users obtained from the activity-based sampling is higher than the number of users from the random sampling. For each day, a higher number

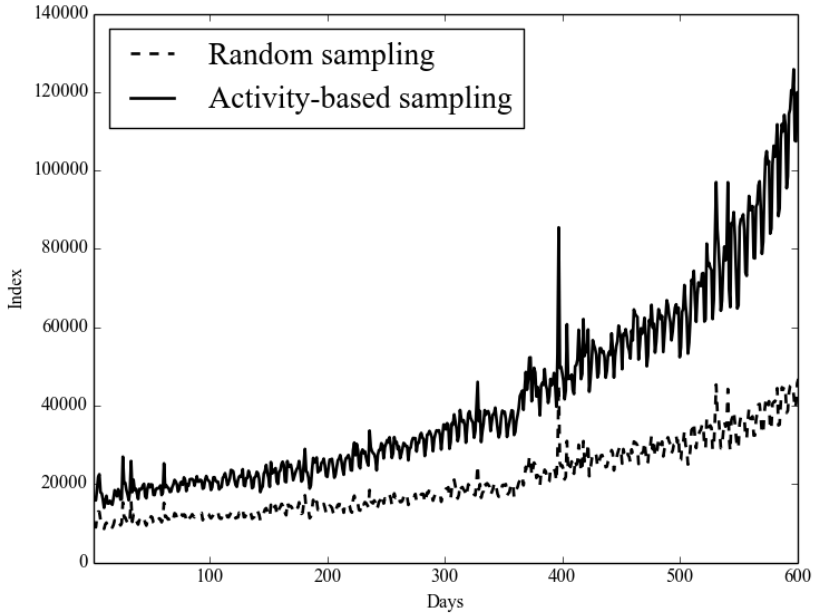
of users were active for the activity-based sampling compared to the random users. In user-centric models, the number of available active users plays an important role in providing interesting patterns for targeted problems [101, 102]. Figure 4.4 also shows the histogram of daily number of tweets and daily number of users to better indicate the coverage of the both sampling approaches.

The statistics of historical tweets and users presented in Table 4.1 indicated that the activity-based sampling compared to the random has better coverage in terms of number of tweets and users. In fact, One of the key question of this study is how to efficiently capture historical tweets which then is applied for content-based and user-centric temporal models. Content-based models are challenged with the number of tweets available on daily basis and in user-centric approaches, the number of available active users plays an important role.

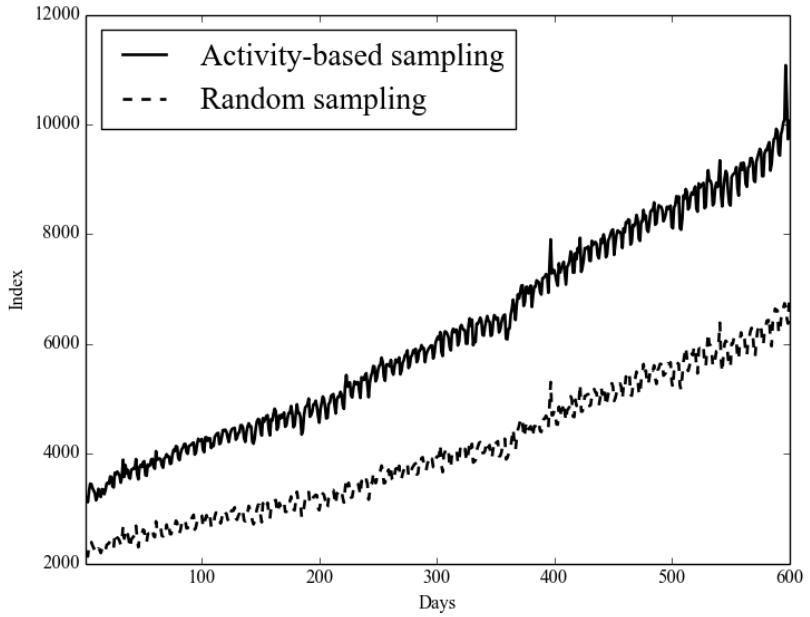
TABLE 4.1: Statistics on the size of users and posts observed on a daily basis for both sampling approaches.

	Daily users		Daily tweets	
	Activity-based	Random	Activity-based	Random
MIN	3,116	2,128	13,952	8,555
STD	1,987	1,326	24,061	9,135
AVG	6,328	4,160	41,568	20,591
MAX	11,077	7,131	125,782	45,352

In more details we are also interested to know how many tweets each individual has posted. In general, the REST API has a limitation of providing only 3,200 (or slightly higher) number of tweets of a specific user. However, if the targeted user did not post more than 3,200 tweets, we can retrieve entire timelines of the selected user. Figure 4.5 show the distribution of overall posts between users for the random and activity-based sampling. The Figures capture very interesting pattern that most of the users (5,350) in the activity-based approach are active and have more than 3,000 number of tweets. Surprisingly, many users in the random sampling were not active during the selected period of time. They were mostly active only during the time data were collected and had no contributions in the past. More than 3,000 users from the random users had

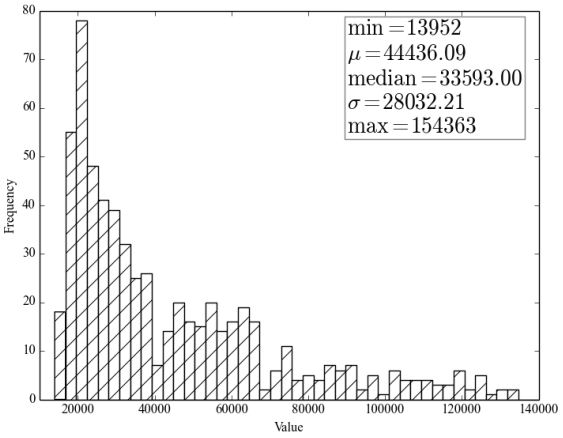


(A) Daily number of tweets. The x-axis shows the days and the y-axis presents the number of tweets.

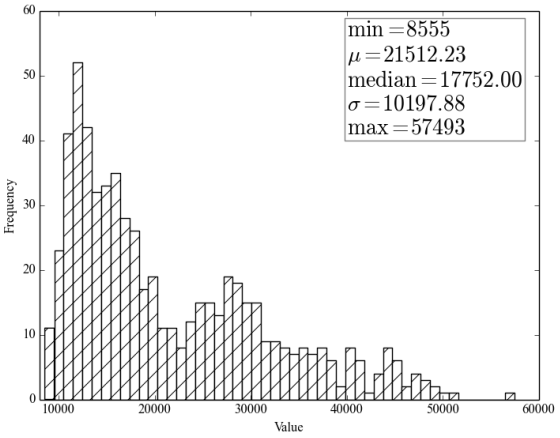


(B) Daily number of users. The x-axis shows the days and the y-axis presents the number of users.

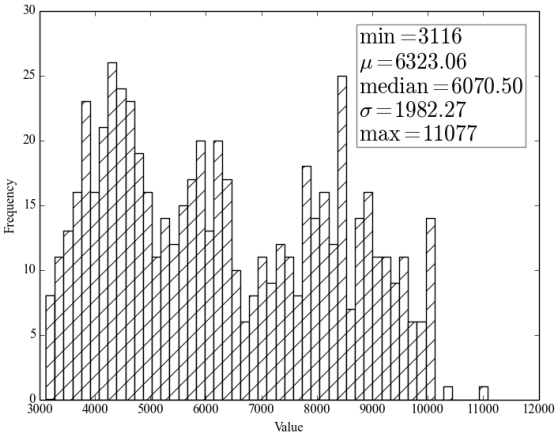
FIGURE 4.3: Daily number of tweets (a) and active users (b) captured from activity-based and random datasets.



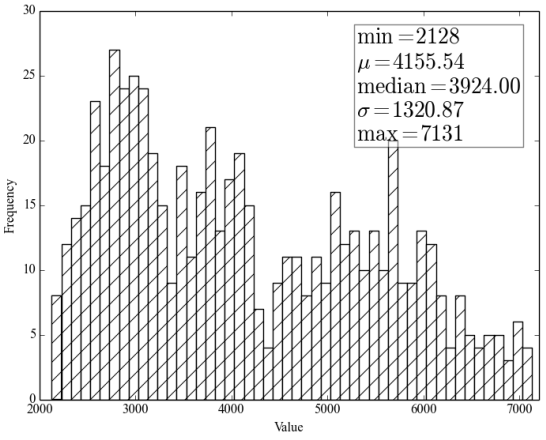
(A) Activity-based (Number of tweets)



(B) Random sampling (Number of tweets)



(C) Activity-based (Number of users)



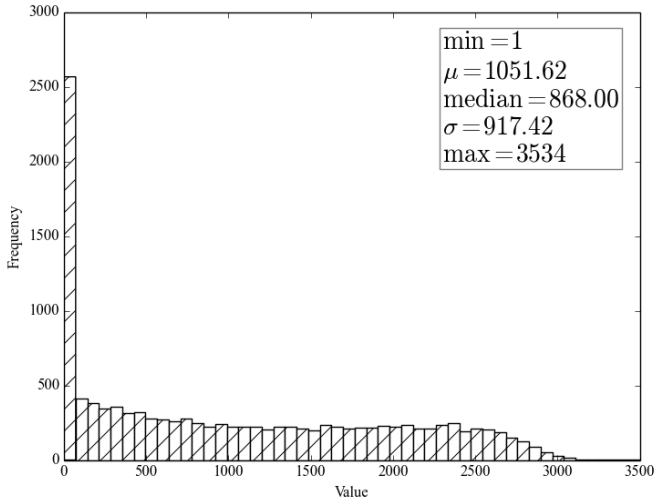
(D) Random sampling (Number of users)

FIGURE 4.4: Histogram of daily number of tweets and active users captured from the activity-based sampling and the random sampling. The x-axises show the daily number and the y-axises present the frequencies.

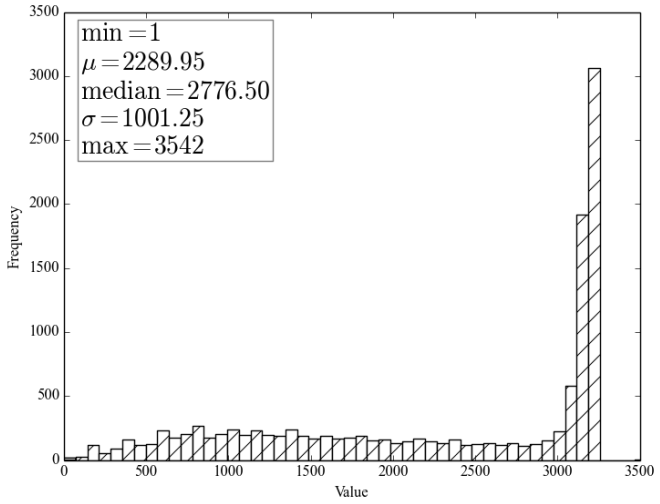
less than 100 tweets during the past. The selected users in the random sampling do not have long time contribution in posting tweets. They were mostly active during data collection, therefore, the number of historical tweets were not significant.

4.3.3 Comparing Activity Gaps

We also investigate the presence of user activity over time, which is the key element in user-centric approaches. Models directly working with user streams are prone to vast amounts of missing opinions. The absent data are related to the errors occurring



(A) Random.



(B) Activity-based.

FIGURE 4.5: Distribution of overall posts between users.

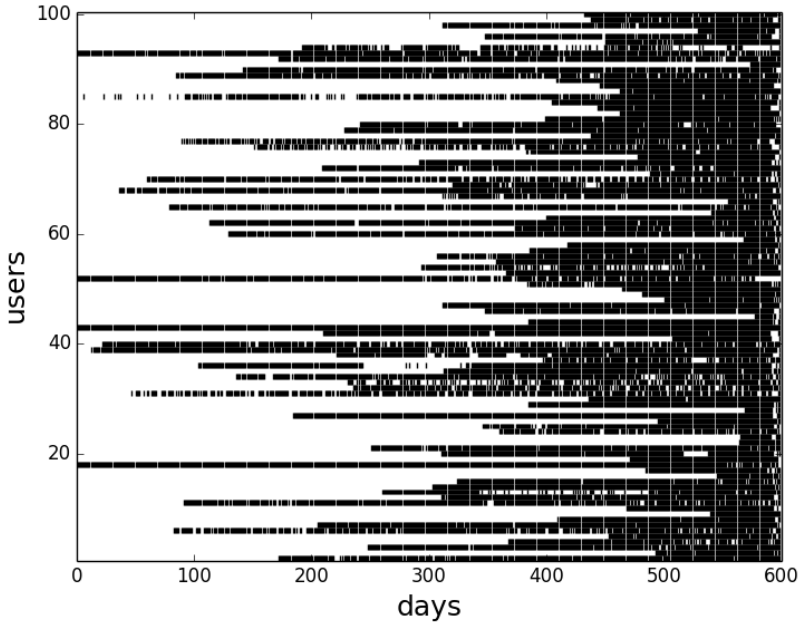
during data collection or simply users are not active during a specific period of time. Although activity gaps are inevitable, it is crucial to retrieve the most active users while avoiding activity gaps in their timelines. While random sampling ignores the activity of the selected users during the past, the activity-based sampling selects users based on their historical timelines. Figure 4.6 shows the daily activity of the 100 most active users during 632 days for the activity-based and random sampling respectively. In this figure, the indexes of the users (y axis) were plotted against the period of time (x axis). The vertical black bar indicates a user has at least one tweet in that

specific days where the white space shows the absence of the user. In fact, the figure indicates the activity of each user over the consideration period of time. Although the top 100 active users, who posted the highest number of tweets, were selected from both approaches; the activity gaps in the random sampling is inevitable. It can be due to the selection of users based on their activities in the streaming time rather than their historical contributions. However, from the Figure 4.6, we can observe that the activity-based approach significantly reduces the absent data, which are more applicable in user-centric approaches.

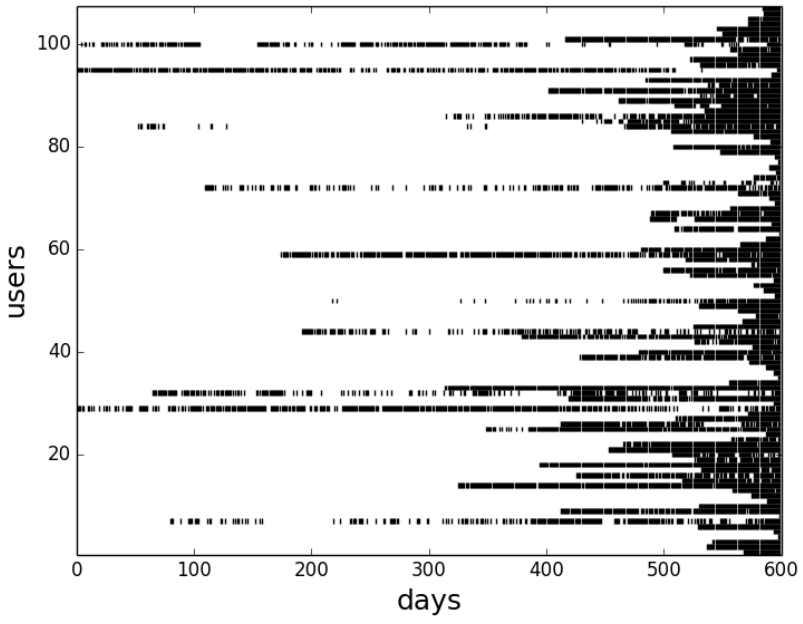
4.3.4 Comparing Credibility

We evaluated the credibility of the datasets in prediction models. The predictability of the content extracted from active users were compared with the content retrieved from the random users in two models: the content-based and user-centric approaches. As discussed before, both models are temporal classification models with different document generation approaches. The classifier is linearSVC, which is the implementation of liblinear [103]. LinearSVC is faster compared with LinearSVM, since kernel transforms are not used and it scales better for large datasets in a linear classification problem. The evaluation was processed by calculating the Macro-averaged F1-score and using rolling origin [104] as the common method for training and evaluating the performance of the model for series observations. In this approach, the training set is the first i (80% of the dataset) and it is tested on the $i+1th$ document. In the second iteration, the training set is moved one document forward (the first $i+1$) and it is tested on the $i+2th$ document. This process is continued until all the test data is classified.

The experiments are conducted in which the content is applied to predict crime trends with different lags. In this regard, document x_i which has been generated at time t_i,



(A) Rastergram of daily activity by 100 most active users for the activity-based sampling.



(B) Rastergram of daily activity by 100 most active users for the random sampling.

FIGURE 4.6: Rastergram of daily activity.

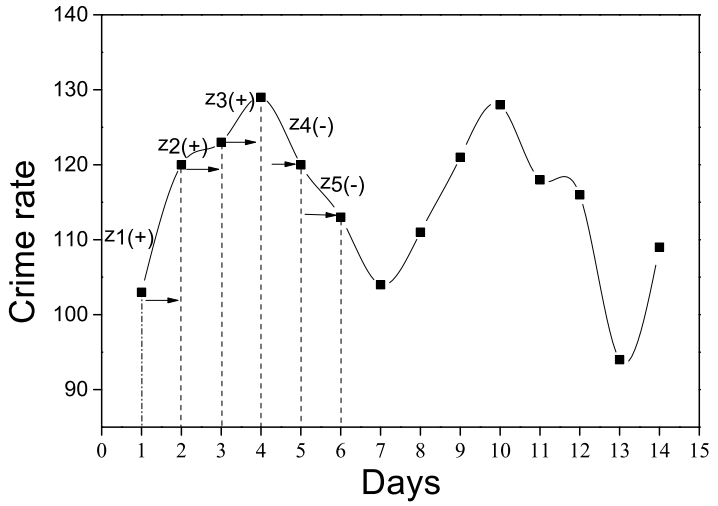
is labeled with crime trend l_i with different lags (see Equation 3.6 in Chapter 3). The lag does not stand for a week or a day, it is a window of time in which crime rate directions are captured. As an example, if lag = 1 ($\Delta r = 1$), each document is labeled with the direction of the crime trend in a day later. In each different lags, the classifier is fed with the generated training data separately. For instance, Figure 4.7 shows the crime trend of BATTERY between a period of 14 days and the generated labels (either +1 or -1) for lag = 1 (Figure 4.7a) and lag = 2 (Figure 4.7b) respectively. In the case of these two different lags, documents are labeled as presented in Table 4.2, where x_1 is a document aggregated at time $t(1)$ and l_1 is its assigned label. The performance of the classifier in lag = 1 and 2 are evaluated separately.

TABLE 4.2: Labeling approach for lag = 1 and lag = 2.

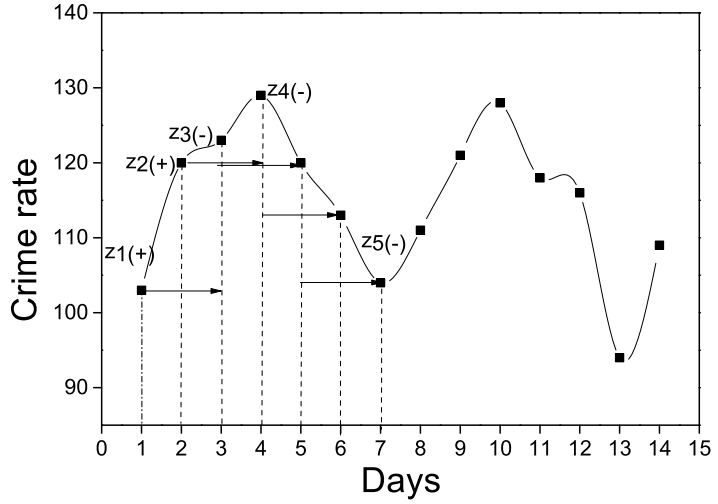
Lag = 1	Lag = 2
$x_1 \rightarrow l_1 : sgn\|y_2 - y_1\| = +1$	$x_1 \rightarrow l_1 : sgn\|y_3 - y_1\| = +1$
$x_2 \rightarrow l_2 : sgn\|y_3 - y_2\| = +1$	$x_2 \rightarrow l_2 : sgn\|y_4 - y_2\| = +1$
$x_3 \rightarrow l_3 : sgn\|y_4 - y_3\| = +1$	$x_3 \rightarrow l_3 : sgn\|y_5 - y_3\| = -1$
$x_4 \rightarrow l_4 : sgn\|y_5 - y_4\| = -1$	$x_4 \rightarrow l_4 : sgn\|y_6 - y_4\| = -1$
$x_5 \rightarrow l_5 : sgn\|y_6 - y_5\| = -1$	$x_5 \rightarrow l_5 : sgn\|y_7 - y_5\| = -1$
$x_6 \rightarrow l_6 : sgn\|y_7 - y_6\| = -1$	$x_6 \rightarrow l_6 : sgn\|y_8 - y_6\| = -1$
$x_7 \rightarrow l_7 : sgn\|y_8 - y_7\| = +1$	$x_7 \rightarrow l_7 : sgn\|y_9 - y_7\| = +1$

4.3.4.1 Prediction Performance: Content-based Model

As discussed before, the content-based model is based on generating documents from aggregating content as bulk with regards to the tweets' timestamps. For pre-processing, we removed stop-word as well as low and high frequent words. We also employed chi-squared for feature selection. The documents were examined with binary and tf-idf representations, however, the best results were achieved using binary



(A) Lag = 1



(B) Lag = 2

FIGURE 4.7: The crime rates of Battery during 14 days and the labeling approach based on (a) lag = 1 and (b) lag = 2.

representation. Table 4.3 illustrates the F-measure of the prediction for the content-based model where both, the content of active and random users, were employed over 7 lags. The highlighted results indicate which content (activity-based or random) is more credible with respect to a specific lag. As an example, the content of the activity-based sampling is more predictive (F-measure = 0.67) compared to the content of random users (F-measure = 0.65) for NARCOTICS when the lag =7. The results demonstrate that the performance of the activity-based sampling for most of the lags are higher than the random sampling. The content of activity-based sampling has higher predictability in accumulation of all crime types. The predictability of the activity-based

content is 27% higher than the random sampling, which indicates the effectiveness of the activity-based sampling for the content-based model where the objective is to predict the directions of indexes. However, in some cases such as BURGLARY and PUBLIC VIOLATION, the difference between the predictability of the two datasets is not considerable. Overall, the results indicate that the proposed activity-based sampling generates more predictive content for ALL and most of the crime types, such as BATTERY, NARCOTICS, and PROSTITUTION, with F-measure up to 0.86.

4.3.4.2 Prediction Performance: User-centric Model

The same sets of experiments were conducted for the user-centric model in which the intention is to leverage individual timelines for document generation. The documents were presented with normalized sentiment scores as discussed in Chapter 3. We examined the credibility of documents with different labels. The results were presented in Figure 4.8 for all users as well as the top 500 users (Figure 4.9). From the results we can observe that in most cases (lags), the content obtained from the activity-based has a higher predictability compared with the random sampling. In the best case, "All" crime with $lag = 6$, the activity-based sampling has achieved an F-measure up to 0.85, which is 35% higher than the random sampling. Overall, the content of active users were shown to be more credible for the proposed user-centric model, which can be the result of having fewer activity gaps compared with the random sampling. In fact, according to the results, the importance of the activity-based sampling for the user-centric model is more significant compared with the content-based model.

Tables 4.4 and 4.5 also show the result of prediction over seven different lags when positive and negative sentiments are employed separately. The results of using positive sentiment in Table 4.4 indicate that in most cases, the content of the activity-based

Table 4.3: The prediction performance for content-based over 7 lags.

Activity-based

	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5	Lag = 6	Lag = 7
Narcotics	**0.51**	**0.54**	0.52	0.53	**0.58**	0.53	**0.67**
Interference	**0.66**	**0.57**	**0.56**	**0.54**	**0.62**	0.52	**0.64**
Deceptive	**0.65**	0.52	**0.57**	**0.64**	0.65	0.62	0.51
Criminal tresspass	**0.64**	**0.57**	**0.58**	**0.62**	**0.57**	**0.64**	**0.56**
Criminal damage	**0.43**	0.6	**0.7**	0.65	0.6	0.56	**0.54**
Sexual Assault	**0.55**	**0.67**	**0.55**	**0.57**	0.52	0.51	**0.54**
Burglary	0.52	**0.58**	**0.56**	**0.56**	0.56	0.54	0.52
Battery	**0.61**	**0.7**	0.62	**0.72**	0.67	0.66	**0.6**
Assault	0.46	0.47	**0.57**	0.52	0.5	**0.54**	**0.56**
ChildrenOffense	**0.62**	**0.58**	**0.67**	**0.62**	0.57	**0.59**	**0.62**
Prostitution	0.57	0.59	**0.7**	**0.68**	0.68	0.58	**0.68**
PublicViolation	**0.46**	**0.51**	**0.47**	**0.51**	**0.55**	**0.55**	**0.53**
Robbery	**0.55**	**0.56**	0.47	0.55	**0.52**	**0.52**	**0.56**
SexOffencse	**0.58**	**0.57**	**0.57**	**0.62**	0.61	**0.59**	**0.61**
Theft	**0.65**	0.55	0.52	**0.6**	**0.62**	0.58	**0.62**
WeaponViolation	**0.64**	0.52	0.51	**0.62**	**0.61**	**0.54**	**0.55**
All	**0.77**	**0.74**	**0.7**	**0.86**	**0.76**	**0.7**	**0.73**

Random

	Lag = 1	Lag = 2	Lag = 3	Lag = 4	Lag = 5	Lag = 6	Lag = 7
Narcotics	0.5	0.51	**0.57**	**0.55**	0.55	**0.55**	0.65
Interference	0.64	0.54	0.52	0.51	0.53	**0.54**	0.55
Deceptive	0.63	**0.56**	0.55	0.55	**0.68**	**0.67**	**0.60**
Criminal tresspass	0.51	0.53	0.57	0.56	0.55	0.57	0.52
Criminal damage	0.42	**0.62**	0.69	**0.68**	**0.65**	0.56	0.51
SexualAssualt	0.52	0.51	0.54	0.55	0.52	**0.52**	0.51
Burglary	**0.53**	0.5	0.52	0.53	**0.57**	0.54	0.52
Battery	0.46	0.67	**0.65**	0.71	**0.7**	**0.7**	0.57
Assault	0.46	**0.54**	0.56	0.54	**0.53**	0.52	0.55
ChildrenOffense	0.58	0.57	0.61	0.60	**0.61**	0.54	0.57
Prostitution	**0.62**	**0.62**	0.67	0.67	0.68	**0.64**	0.54
PublicViolation	0.44	0.47	0.46	0.46	0.47	0.53	0.49
Robbery	0.5	0.54	**0.51**	**0.56**	0.5	0.49	0.44
SexOffense	0.54	0.55	0.55	0.57	0.60	0.54	0.52
Theft	0.57	**0.57**	**0.56**	0.53	0.61	0.58	0.55
WeaponViolation	0.57	**0.58**	**0.52**	0.57	0.52	0.52	0.53
All	0.5	0.59	0.6	0.59	0.54	0.58	0.61

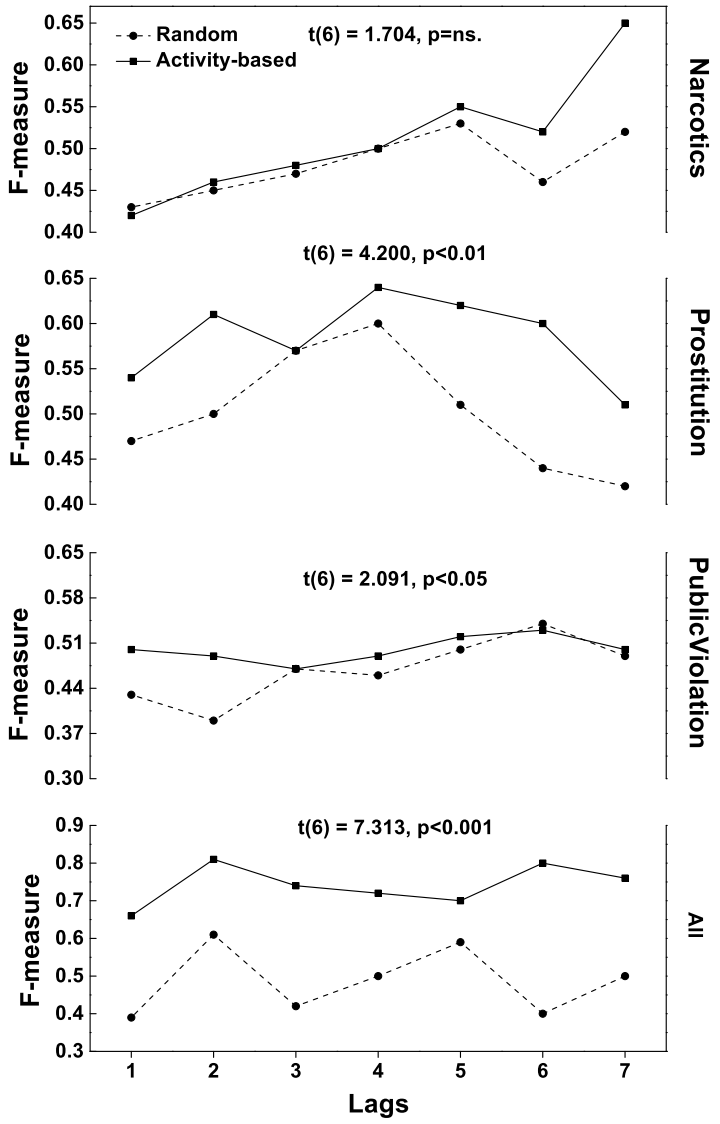


FIGURE 4.8: Predictability for user-centric approach over 7 lags for "all users". "All" is the overall crime rates.

approach achieved more predictability compared to the content of random users. The same pattern was observed when negative sentiment is employed (Table 4.5). The highlighted results show in which case the content of the activity-based sampling is higher than the content of random users with respect to a specific lag. However, the best results over seven lags for DECEPTIVE PRATICE, CRIMINAL TRESSPASS, CHILDREN OFFENSE and PROSTITUTION were presented in Figure 4.10 and for all crimes in Table 4.6. Overall, the content of users sampled based on their activities were shown to

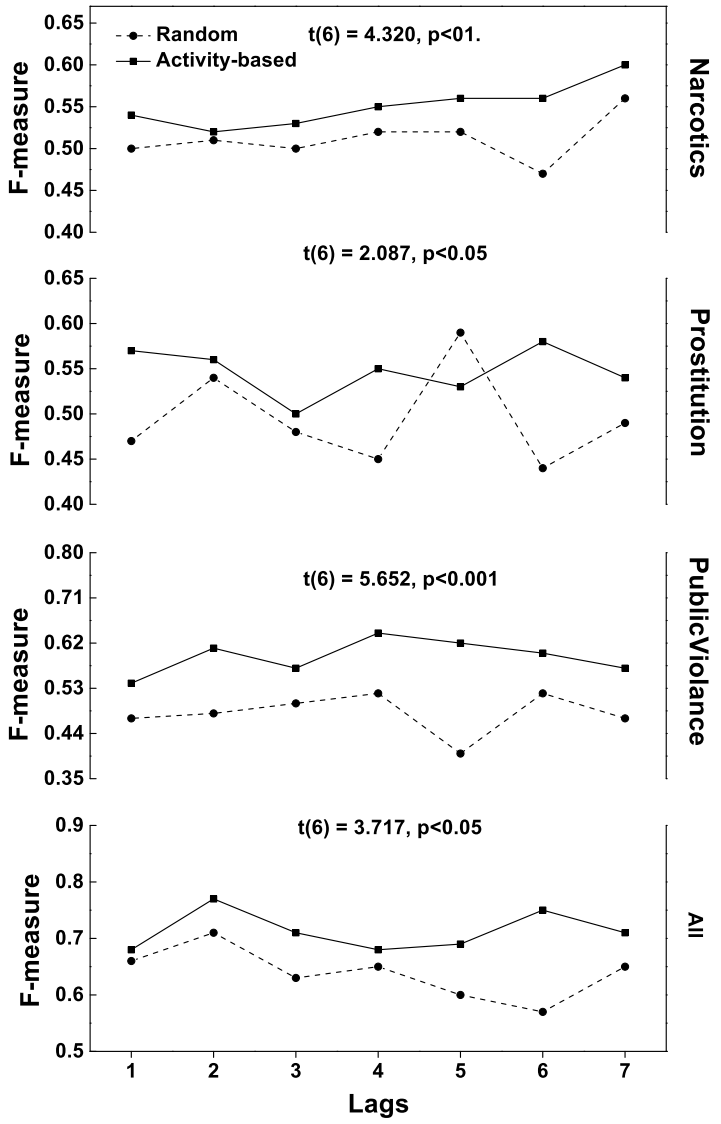


FIGURE 4.9: Predictability for user-centric approach over 7 lags for "Top 500 users". "All" is the overall crime rates.

be more credible for the proposed user-centric model, which can be the result of having less sparsity in users' activities compared with the random sampling. In fact, the activity-based sampling captures users with fewer activity gaps, therefore, the availability of content over time significantly effects the predictability as shown in this experiment.

Table 4.4: Macro F-measure of the prediction performance for active and random users based on positiveness.

Random

	$\Delta r = 1$	$\Delta r = 2$	$\Delta r = 3$	$\Delta r = 4$	$\Delta r = 5$	$\Delta r = 6$	$\Delta r = 7$
Narcotics	0.42	0.47	0.51	0.52	**0.56**	0.49	0.58
Interference	0.41	0.48	0.52	0.51	0.43	0.49	0.44
Deceptive	0.44	0.48	0.5	**0.59**	**0.62**	0.53	**0.59**
Criminal tresspass	0.48	0.5	0.51	0.52	**0.59**	0.47	0.52
Criminal damage	0.36	**0.62**	0.56	**0.56**	0.6	0.51	0.47
SexualAssualt	0.43	0.46	0.52	**0.56**	0.5	**0.6**	0.47
Burglary	0.48	0.52	0.48	0.51	0.5	**0.6**	0.51
Battery	0.45	0.55	0.6	0.6	**0.58**	**0.6**	0.55
Assualt	0.4	0.46	0.48	0.47	**0.58**	0.48	0.51
ChildrenOffense	0.43	**0.52**	**0.54**	0.5	**0.56**	0.52	0.4
Prostitution	0.55	0.57	**0.62**	0.63	0.55	0.52	0.52
PublicViolation	0.37	0.38	0.43	0.46	0.49	0.5	0.42
Roberry	**0.46**	**0.54**	0.47	**0.53**	0.44	0.43	0.5
SexOffense	0.39	0.5	0.5	0.49	0.54	0.55	0.52
Theft	0.46	0.49	0.51	**0.6**	0.56	**0.58**	0.53
WeaponViolation	0.42	0.5	0.48	0.51	0.53	0.47	0.42
All	0.65	0.77	0.67	0.71	**0.76**	0.65	0.7

Activity-based

	$\Delta r = 1$	$\Delta r = 2$	$\Delta r = 3$	$\Delta r = 4$	$\Delta r = 5$	$\Delta r = 6$	$\Delta r = 7$
Narcotics	**0.55**	**0.59**	**0.66**	**0.55**	0.55	**0.56**	**0.65**
Interference	**0.58**	**0.53**	**0.62**	**0.55**	**0.61**	**0.52**	**0.51**
Deceptive	**0.59**	**0.58**	**0.52**	0.55	0.55	**0.58**	**0.55**
Criminal tresspass	**0.55**	**0.59**	**0.67**	0.52	0.58	**0.53**	**0.57**
Criminal damage	**0.57**	0.57	**0.66**	0.55	**0.64**	**0.64**	**0.58**
SexualAssualt	**0.58**	**0.57**	**0.59**	0.55	**0.62**	**0.67**	**0.51**
Burglary	**0.59**	**0.57**	**0.62**	**0.52**	**0.59**	0.53	**0.59**
Battery	**0.58**	**0.62**	0.6	0.6	0.6	0.58	**0.6**
Assualt	**0.65**	**0.64**	**0.66**	**0.59**	0.54	0.54	**0.53**
ChildrenOffense	**0.55**	0.51	0.5	**0.63**	0.52	**0.65**	**0.58**
Prostitution	**0.67**	**0.61**	0.57	**0.71**	**0.62**	**0.6**	**0.54**
PublicViolation	**0.55**	**0.64**	**0.57**	**0.6**	**0.52**	**0.54**	**0.57**
Roberry	0.44	0.51	**0.54**	0.48	**0.55**	**0.47**	**0.52**
SexOffense	**0.55**	**0.57**	**0.59**	**0.59**	**0.66**	**0.58**	**0.53**
Theft	**0.55**	**0.53**	**0.52**	0.52	**0.63**	0.51	**0.59**
WeaponViolation	**0.52**	**0.55**	**0.5**	**0.54**	**0.55**	**0.58**	**0.53**
All	**0.68**	**0.81**	**0.74**	**0.72**	0.7	**0.8**	**0.76**

Table 4.5: Macro F-measure of the prediction performance for active and random users based on negativeness.

Random

	$\Delta r = 1$	$\Delta r = 2$	$\Delta r = 3$	$\Delta r = 4$	$\Delta r = 5$	$\Delta r = 6$	$\Delta r = 7$
Narcotics	0.4	0.48	0.49	0.47	0.51	0.52	0.6
Interference	0.4	0.41	0.44	0.48	0.44	0.48	0.41
Deceptive	0.44	0.5	0.45	**0.59**	0.6	0.53	0.46
Criminal tresspass	0.38	0.52	0.5	**0.59**	**0.49**	0.51	0.57
Criminal damage	**0.51**	0.47	0.53	0.59	**0.57**	0.47	**0.55**
SexualAssualt	0.4	0.47	0.57	0.47	0.44	0.51	**0.49**
Burglary	0.38	0.42	0.46	0.54	0.49	0.53	0.47
Battery	**0.47**	0.58	0.53	0.6	0.52	**0.6**	0.5
Assualt	0.43	0.44	0.5	0.48	0.42	0.51	0.52
ChildrenOffense	0.4	0.42	**0.58**	0.44	0.55	0.49	**0.56**
Prostitution	0.46	0.47	**0.51**	**0.59**	**0.57**	0.58	0.48
PublicViolation	0.43	0.39	0.47	0.46	**0.5**	**0.54**	**0.49**
Roberry	0.35	0.43	0.43	0.45	0.49	0.48	**0.54**
SexOffense	0.44	0.43	0.47	0.53	0.51	0.5	0.53
Theft	0.43	0.46	0.48	**0.59**	0.46	0.5	**0.49**
WeaponViolation	0.48	0.44	0.37	0.43	0.47	0.46	0.45
All	0.65	0.76	0.66	0.7	**0.78**	0.71	**0.72**

Activity-based

	$\Delta r = 1$	$\Delta r = 2$	$\Delta r = 3$	$\Delta r = 4$	$\Delta r = 5$	$\Delta r = 6$	$\Delta r = 7$
Narcotics	**0.52**	**0.56**	**0.5**	**0.55**	0.51	**0.56**	**0.61**
Interference	**0.49**	**0.53**	**0.48**	0.48	**0.51**	**0.55**	**0.47**
Deceptive	**0.6**	**0.62**	**0.69**	0.55	**0.61**	**0.6**	**0.54**
Criminal tresspass	**0.49**	**0.56**	**0.55**	0.56	0.48	**0.59**	**0.6**
Criminal damage	0.49	**0.58**	0.53	**0.68**	0.54	**0.55**	0.53
SexualAssualt	**0.44**	**0.48**	0.57	**0.59**	**0.64**	**0.54**	0.47
Burglary	**0.57**	**0.61**	**0.55**	**0.67**	**0.66**	**0.56**	**0.55**
Battery	0.44	**0.64**	0.53	0.6	**0.58**	0.56	**0.56**
Assualt	**0.48**	**0.5**	**0.62**	**0.53**	**0.62**	**0.56**	**0.6**
ChildrenOffense	**0.48**	**0.52**	0.52	**0.67**	**0.68**	0.49	0.49
Prostitution	**0.49**	**0.49**	0.48	0.55	0.54	**0.68**	**0.51**
PublicViolation	**0.53**	**0.61**	**0.59**	0.47	0.47	0.53	0.45
Roberry	**0.46**	**0.66**	**0.55**	**0.62**	**0.51**	0.48	0.49
SexOffense	**0.46**	**0.46**	**0.5**	0.53	**0.63**	**0.67**	**0.54**
Theft	**0.54**	**0.63**	**0.52**	**0.53**	**0.53**	**0.54**	0.43
WeaponViolation	**0.53**	**0.52**	**0.48**	**0.5**	**0.56**	**0.56**	**0.51**
All	0.65	**0.83**	**0.76**	**0.75**	0.75	**0.81**	0.69

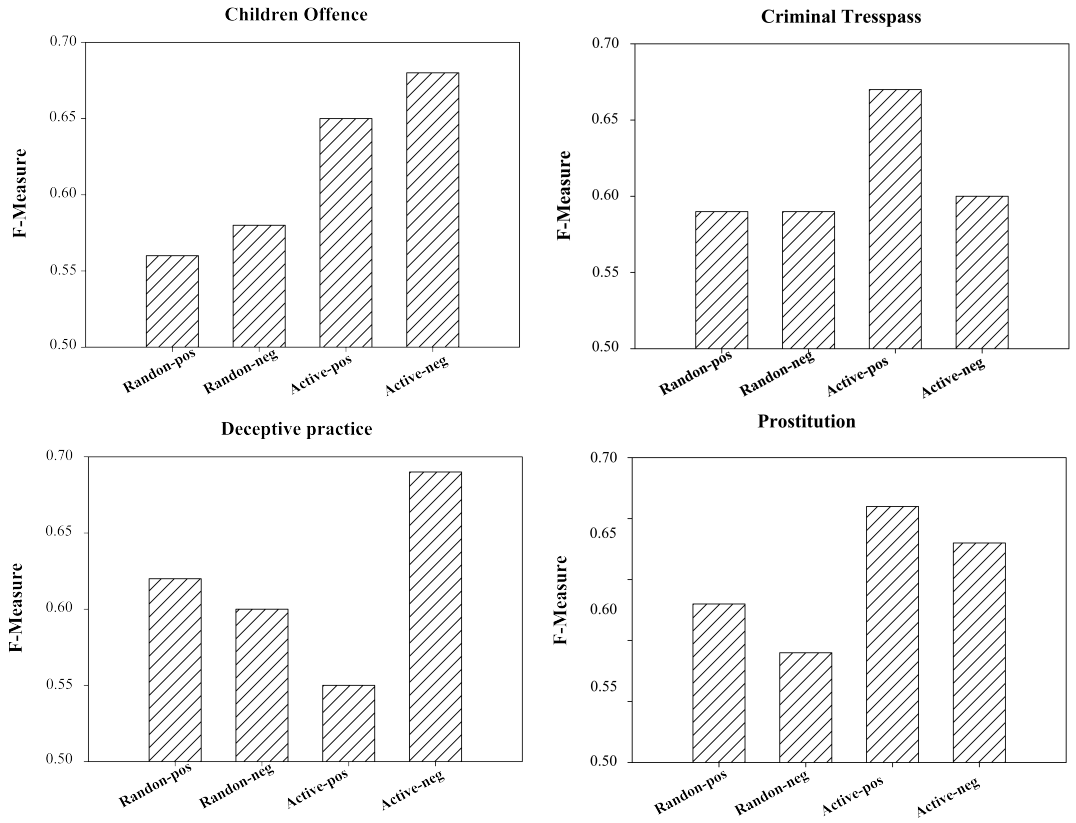


FIGURE 4.10: The best results obtained by the random and the activity-based approaches for both positive and negative sentiments.

TABLE 4.6: The best results obtained by the random and the activity-based approaches for both positive and negative sentiments.

Crime type	Random_pos	Random_neg	Activity-based_pos	User filtering_neg
Narcotics	0.58	0.6	**0.66**	0.61
Interference	0.52	0.48	**0.59**	0.55
Deceptive	0.62	0.6	0.55	**0.69**
Criminal tresspass	0.59	0.59	**0.67**	0.6
Criminal damage	0.62	0.57	0.66	**0.68**
SexualAssualt	0.6	0.57	**0.67**	0.64
Burglary	0.6	0.54	0.62	**0.67**
Battery	0.6	0.6	0.62	**0.64**
Assualt	0.58	0.52	**0.66**	0.62
ChildrenOffense	0.56	0.58	0.65	**0.68**
Prostitution	0.63	0.59	**0.71**	0.68
PublicViolation	0.5	0.54	**0.64**	0.61
Roberry	0.54	0.54	0.55	**0.66**
SexOffense	0.55	0.53	0.66	**0.67**
Theft	0.6	0.59	**0.63**	**0.63**
WeaponViolation	0.53	0.48	**0.58**	0.56
All	0.77	0.78	0.81	**0.83**

4.4 Conclusions

Identifying credible sources of content or users are important in many research problems aiming to drive a meaningful conclusion from the source of information. Performance of the prediction models can be degraded from the missing data or the choice of the collected content. This thesis has argued the importance of the selection of data for the targeted problem. In this chapter, we focused on sampling Twitter users to retrieve their historical tweets for temporal prediction models. We presented an activity-based approach that leverages user profiles to estimate historical activities in the past for the selection of the most active users as opposed to expert users. In this approach, we selected users based on two factors: the number of days a user is active and the average number of user's tweet per day. Both factors were calculated using user profile elements such as "$created\_at$" and "$status\_count$". In addition to the activity-based sampling, we also gathered another set of users by random sampling, which is widely used to collect users for the REST API.

The historical timelines of the selected users were also retrieved using the REST API. The timelines of the collected tweets from both groups of users were limited to our period of time consideration. We compared the primary statistical differences between two datasets in terms of historical timelines and users' activities. Regarding the number of tweets and users, the activity-based approach has better coverage compared to the random samples. We also compared the overall number of tweets for each user. Most of the users were active for the activity-based sampling, and the random users had low activity during the selected period of time. In addition, the activity gap of both sets of users were compared. The results indicate that active users had more contributions in the past, while activity gaps in the random sampling are inevitable. In fact, the activity-based sampling significantly reduces the absent data because users were selected based on their histories. Overall, the activity-based approach identifies users

who are more historically active, whereas in the random sampling high activity gaps are observed.

In addition, we also studied the credibility of the content captured from both datasets in the proposed temporal prediction models. We presented two temporal prediction models (user-centric and content-based) to compare the credibility of the content gathered from the selected users. While, in the content-based model, documents are generated based on historical tweets of all collected users, in the user-centric, documents are created based on individual timelines. Both models were applied to predict the directions of crime rates. The prediction models leverage historical tweets to predict crime rate increases or decreases for the prospective timeframe.

The results of the content-based model indicate that the content of active users is more credible in predicting the trend of interest. In the best case, the results is 27% higher when using the content of active users. Overall, in 16 crime types out of 17, the activity-based approach achieved the best results compared to the random sampling. This is due to the fewer activity gaps observed in the collected tweets of the active users compared to the random users. For temporal content-based models, such as our proposed model, the availability of content over time plays a crucial role. In the user-centric model, the same performance was observed. The content of active users has higher predictability. In fact, the user-centric model relies on the availability of timelines of users, which is highly affected by the activity gaps. As the results indicated, the performance was significantly higher in some cases (PUBLIC VIOLATION, ALL) compared to the random sampling. Overall, the intention of this chapter was to show the importance of a target-oriented data sampling for temporal prediction models. Therefore, in the next Chapter, we employ the proposed data sampling approach to collect tweets for the proposed temporal topic model.

Chapter 5

Temporal Topic Model

5.1 Overview

Topic models are extensively applied in many different purposes such as to indicate similarities between two sets of documents [29], or to visualize a high dimensional documents to a set of well-structured variables for exploration [31]. Nevertheless, with the increasing number of user-generated documents in microblogs, there has been great demand for topic models to drive meaningful patterns. Twitter as one of the most popular online services with more than 500 million Tweets sent each day generates enormous social data where topic models can be used for summarizing text streams.

Latent Dirichlet Allocation (LDA) as a generative model in topic modeling [34] has shown to be effective and highly applicable to convert content to a small set of hidden variables. Documents are represented with a mixture of topics which in turn contain a set of word distributions. In LDA model inputs are bag-of-words representation of documents and output is assigned latent topics to each document in corpus. A topic in LDA is multinomial distribution of words in a dictionary, while a document is multinomial distribution of topics. Figure 5.1 presents the plate notation of LDA. A topic model is learned by giving N number of documents with vocabulary of size $|V|$. In the

learning phase, documents are given as a single group which builds a static vocabulary for inferring topic distributions. However, in temporal modeling, emerged documents are most likely carrying new terms, therefore, predefining a fixed vocabulary is not practical. While topics have proven to have birth and death [6], when extracted from temporal text streams, there is a significant need of having different vocabularies over time. Overall, in topic identification from temporal documents many major variables such as size of vocabulary, number of topics, and topics distributions are reliant on time. Another raised issue is inference of insignificant topics when applying static LDA. Because of having broad range of documents (daily aggregation of conversations over long period of time), topics consist of common words more likely are generated if documents are given as a single batch. In addition, significant social events have high impacts on Twitter content. As an example, in some specific periods of time, high numbers of tweets can be observed due to coincident with some events which carry information related to the events. To tackle the aforementioned issues, we propose a model with the following characteristics:

- A model which can infer emerged topics and measure topics evolution.
- A model with the size not growing in time and being capable of detecting the best representative topics. Vocabulary is updated, and previous unseen terms fade away.
- A model which does not converge in topics after long period of time. Convergence can be prevented by applying LDA on each period of time rather than transferring learning parameters from previous learned model.
- A model which can handle sequential data and can be updated with introducing new documents.

In this thesis, we have also leveraged topics identified from text streams on Twitter to forecast social index changes. The assumption is that social data contain valuable

information about topics of interest and events which may be correlated with the real world problems. In order to investigate the correlation between discussion topics and social problems, we present a temporal topic detection model. In temporal analysis, time plays a crucial role for topic identification. Information is variant over time and a flow of changes can be observed over long term consideration. As an example, Figure 5.2 presents the frequency of top 2000 words (stemmed words) over two consecutive years (2012 and 2013) from our collected dataset. From the figure, we can observe that the frequency of words are variant over time. As an example "gun","energi", and "basketball" were popular in 2012, while "cancer" and "campaign" were more popular terms in 2013. However, some words such as "data" and "develop" are constantly repeated for the two years with similar word frequencies. In generic topic modeling, LDA trains on static vocabulary where information evolution is not addressed. However, temporal topic model is based on dynamic vocabulary to identify emerged topics over time. In this thesis, we introduce a temporal topic model to automatically identify distinct topics and rank them based on their novelty and diversity. Further, they are applied as representative variables for trend prediction. We also address how topics change over time and how new topics emerge in each period of time. In general, we leverage novelty and diversity in topics to capture social trend changes.

This chapter aims to achieve the following objectives:

- to leverage temporal topics derived from Twitter in social trend prediction. In fact, the main aim is mapping hidden variable extracted from daily conversations with the crime changes in the following days.
- to involve time in topic identification in which, emerged topics are captured from text stream over time.
- to efficiently capture topic evolution and contribute novelty and diversity in identifying topics.

While information evolves over time, the idea of topic extraction with time dimension shed a light to be an efficient model to infer the best predictive latent topics to surpass the performance of prediction model.

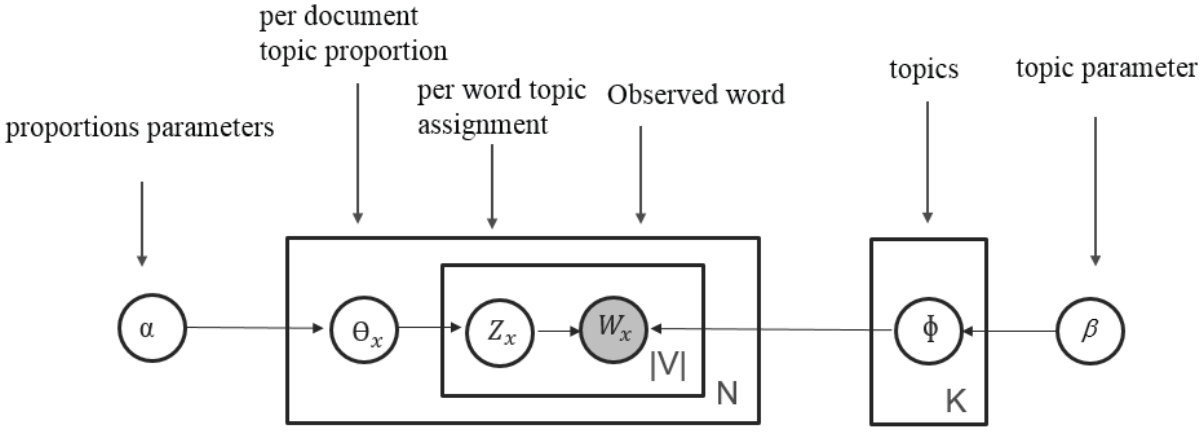


FIGURE 5.1: Graphical representation of LDA. The parameters are as follows: α is the hyper parameter per document topic proportion, in which θ_d is topic distribution inferred for each document, Z_x is inferred from θ_x, in which Z_x is drawn $|V|$ times ($|V|$ is the size of vocabulary), β is hyper parameter for per topic word distribution, and ϕ is word distribution for each topic.

5.2 Temporal Topic Model

In this section, we propose our topical model to infer temporal topics as predictive features for the proposed prediction models. As extensively discussed in the previous section, topics are subject to change [6], when extracted from temporal text streams. Therefore, there is a significant need to have dynamic vocabularies over time to address emerging terms in topic inference and fade away vocabularies which are no longer popular. To tackle the issue of changes of vocabularies in topic inference we develop a model that has four phases: (1) Document segmentation, (2) Topic inference, (3) Topic selection, and (4) Document representation. Figure 5.3 shows the general framework of our temporal topic model.

The generic procedure of the proposed temporal topic model is presented in Algorithm 3. In this model, an LDA model is trained separately for each time slice (partition) and

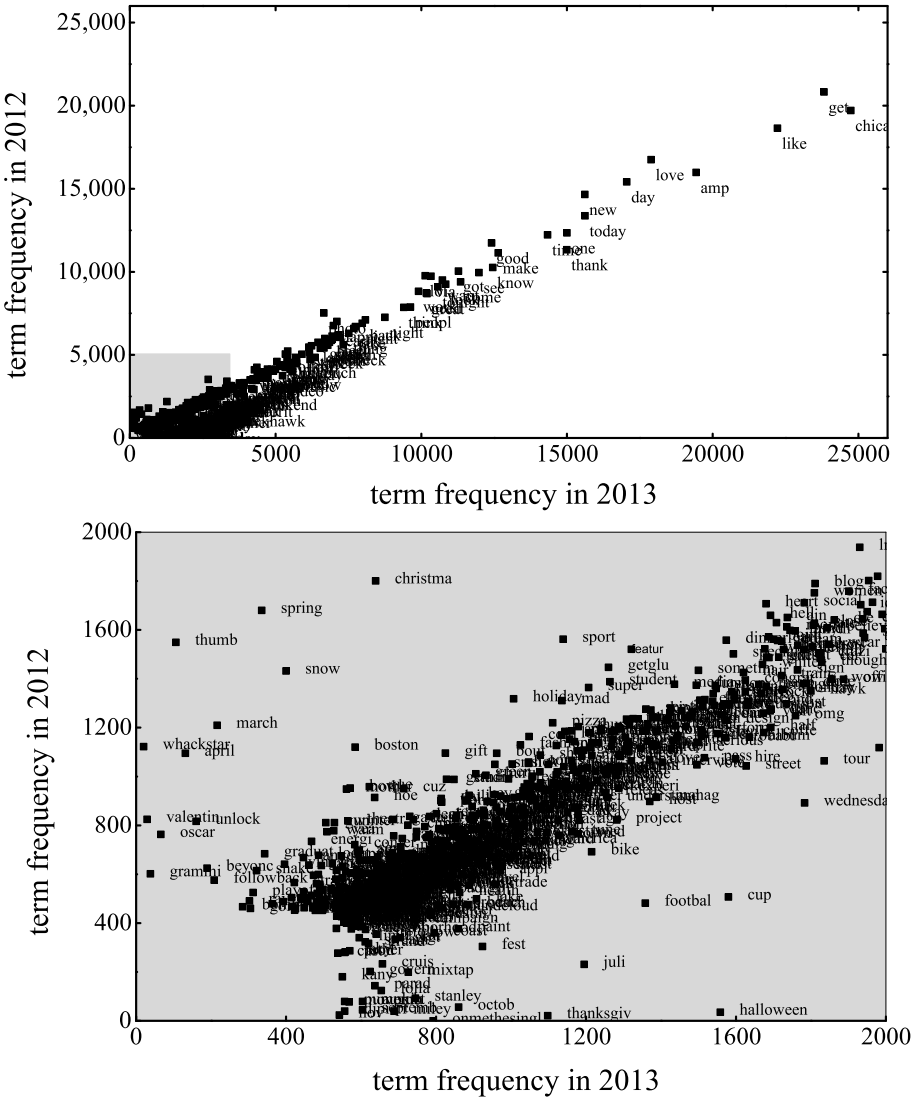


FIGURE 5.2: Word frequency over two different years.

a set of topics is inferred for each partition. A time slice or partition is a unit of time (a month, a year,...) in which documents are classified based on their timestamps. In every iteration, topic similarities between two partitions are estimated. In fact, the approach seeks for the degree of topic similarities between extracted topics in the prospective time slice ($partition_{t+1}$) compared to the current time slice ($partition_t$). The identified topics at $partition_{t+1}$ which are similar to the already detected topics at $partition_t$ are not selected. After selection of the proper topics, topic distributions are inferred for the other unseen documents, based on the LDA models corresponding to the partitions. In fact in Batch LDA, documents are given to LDA and topics

are extracted, while our proposed temporal model divide the documents into different partition. In each partition, the proposed model learns from each time-slice and updates its parameters. The updated learning parameters are used in other slices to infer the temporal topics. In this regard, the vocabulary will be changed. Therefore, we account for new words and different terms frequencies, while Batch LDA defines a static vocabulary for topic inference.

The following subsections discuss the major steps of the proposed temporal topic model.

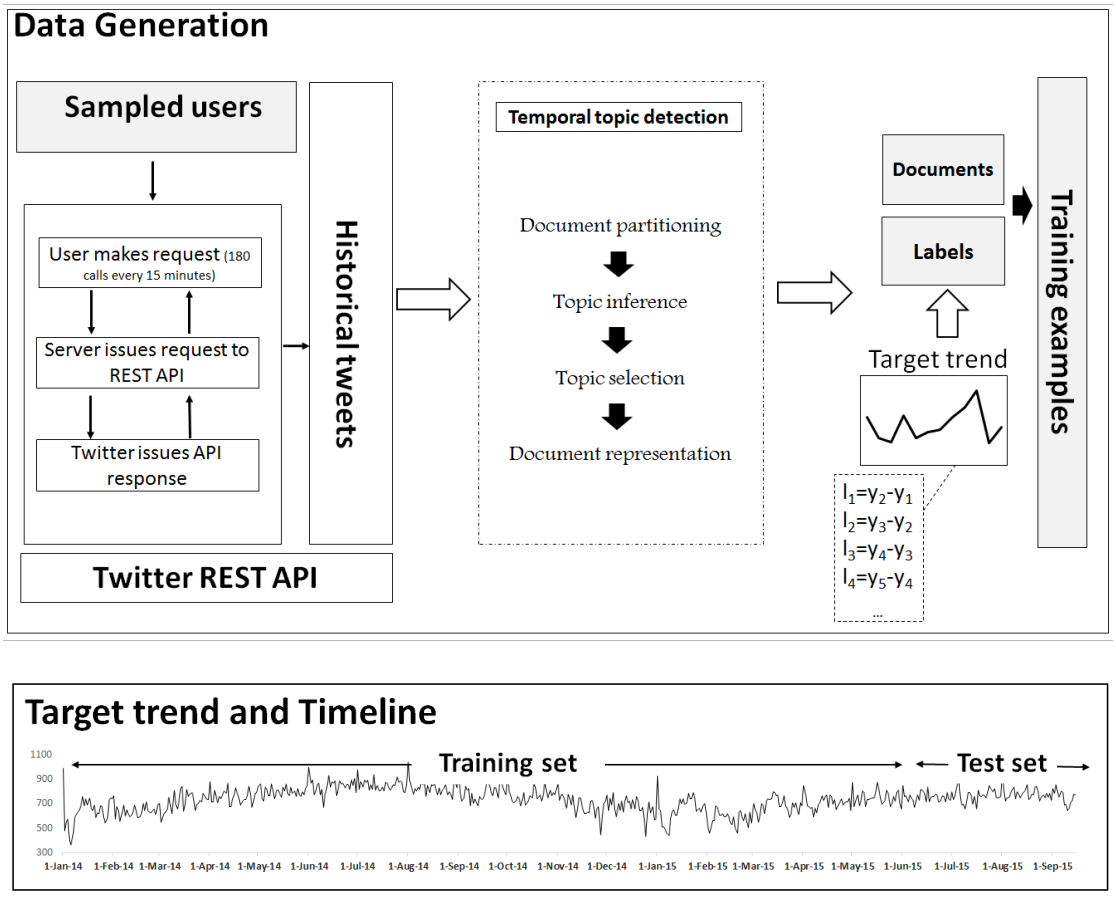


FIGURE 5.3: The framework of the data generation model.

Algorithm 1: Procedures of the temporal topic model

Input: documents X
Output: document-topic matrix

Initialize K and m
Given m, documents are placed into different partitions: $\{(1, ..., m), (m+1, ..., 2m), ...\}$
$u \leftarrow \frac{N-\Delta r}{m}$ //number of partitions
for each partition $j = 1, ..., u$ **do**
    $P_j \leftarrow (x_{[(j-1)m+1]}, ..., x_{[(j-1)m+m]})$
    $Q_j \leftarrow (x_{[(jm+1]}, ..., x_{[jm+m]})$
    $V \leftarrow \bigcup w^{P_j}$ // Regenerate vocab with words occurring in P_j
    $lda_j \leftarrow$ lda (P_j, k^{p_j},V) // estimating LDA parameters based on the training data in P_j
    $X^{temp(P_j)} \leftarrow lda_j[X]$ // inferring topic distribution for all documents based on trained model (lda_j)
    $V \leftarrow \bigcup w^{P_j}$ //Regenerate vocab with words occurring in Q_j
    $lda_{j+1} \leftarrow$ lda (Q_j, k^{Q_j},V) // estimating LDA parameters based on the training data in Q_j
    $X^{temp(Q_j)} \leftarrow lda_{j+1}[X]$ // inferring topic distribution for all documents based on trained model (lda_{j+1})
    for $k := 1$ to K^{Q_j} **do**
        $Sim(T_k^{Q_j}, T_K^{P_j}) \leftarrow \sum_{h=1}^{K^{P_j}} Dist(T_k^{Q_j}, T_h^{P_j})$
        If $Sim(T_k^{Q_j}, T_K^{P_j}) < threshold$ **then**
            Regenerate $X^{temp(Q_j)}$
$X_{n*K} \leftarrow$ accumulation of all $X^{temp(Q_j)}, X^{temp(Q_j)}$;

5.2.1 Document Partitioning

Given all documents X with their timestamps, the documents are placed into different partitions. As an example, if the observation period is 12 months and the size of partition is one month, the documents are partitioned monthly according to their timestamps. Figure 5.5 represents an illustration of temporal partitions for the inference of document-topic and document-term matrices. In each iteration, two different sequential partitions are processed for topic inference. Documents in the first partition are considered as a true representation of data (P) and the second partition (Q) denotes the model (see Figure 5.4). Therefore, partitions are created as follows:

$$P_j = \bigcup_{r=1}^{m} x_{[(j-1)m+r]}$$
$$Q_j = \bigcup_{r=1}^{m} x_{[jm+r]} \tag{5.1}$$

where m is the size of each partition and x denotes the feature vector of a document.

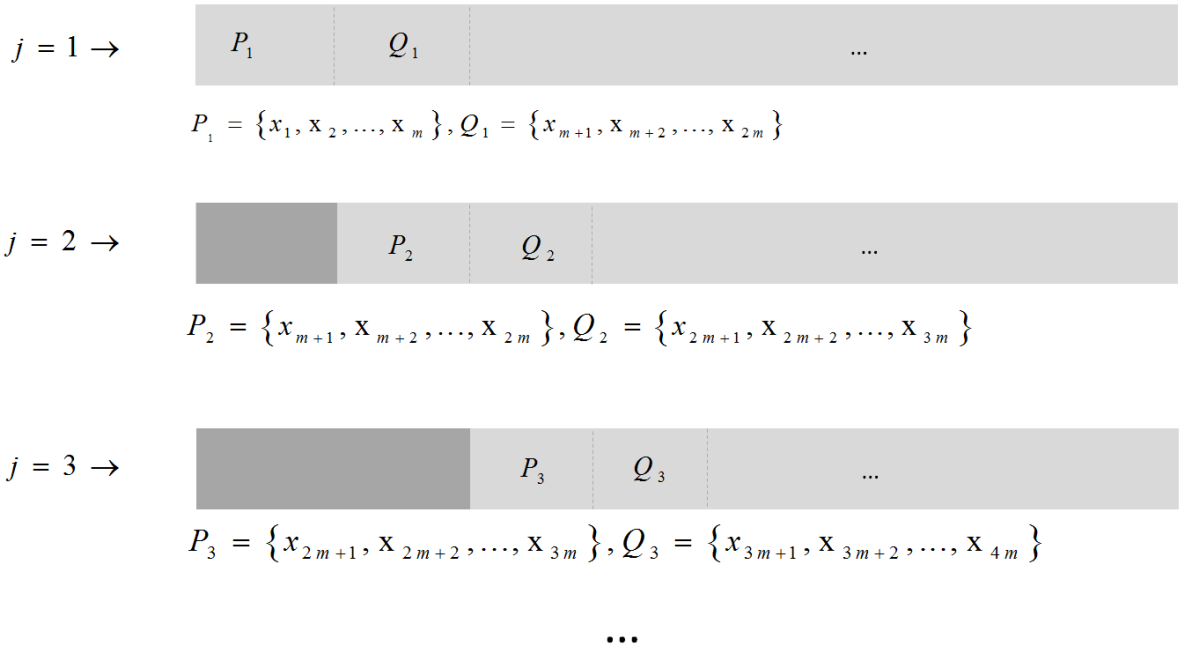


FIGURE 5.4: The general schema of document partitioning.

5.2.2 Topic Inference

In our temporal model, the LDA parameters are inferred as presented in Algorithm 2 and 3. The process of topic inference will be continued by taking the next two partitions (Eq 5.1). In these algorithms, every two partitions, P_i and Q_i, are considered together. First, k^P number of topics is predefined and LDA is processed to estimate parameters on segment P_i (Algorithm 2). Second, at the arrival of a new document for the next partition (Q_j), the LDA parameters are updated using the procedure explained

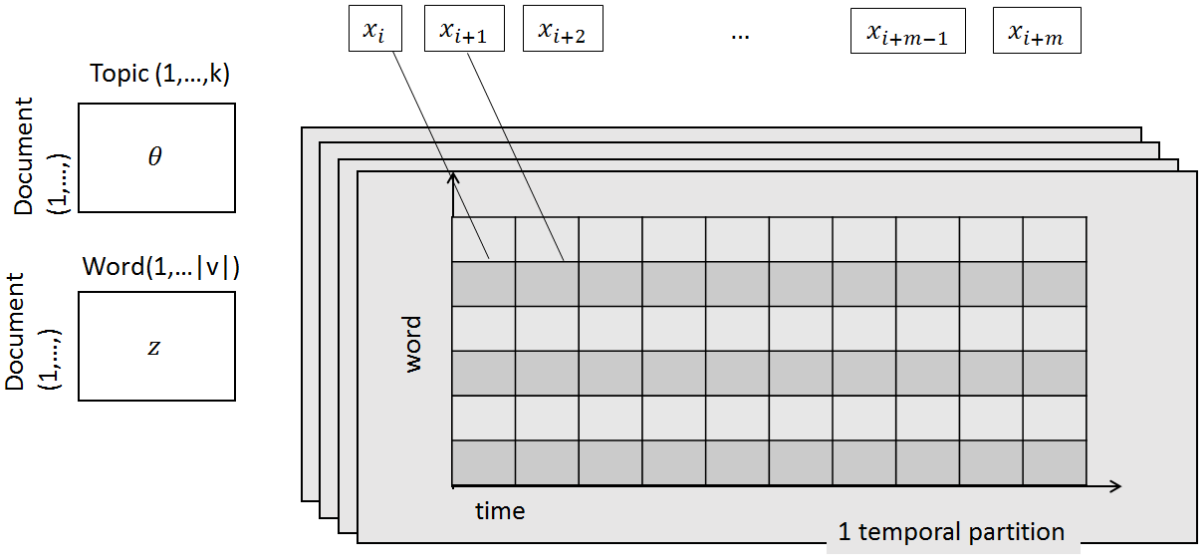


FIGURE 5.5: An illustration of temporal partitions.

Algorithm 2: Generative process of LDA for P_j.

For each topic $1, 2, ..., K_j^P$:
    Draw $\phi_k^{P_j} \sim Dirichlet(\beta^{P_j})$
For each document x_i from P_j:
    Draw a distribution over topics $\theta_{x_i}^{P_j} \sim Dirichlet(\alpha^{P_j})$
    For each word in the document $w \in x_i$:
        Draw a topic z $\sim Multinomial(\theta_{x_i}^{P_j})$
        Draw a word $w \sim Multinomial(\phi_{x_i}^{P_j})$

Algorithm 3: Generative process of LDA for Q_j.

For each topic $1, 2, ..., K_j^Q$:
    Draw $\phi_k^{Q_j} \sim Dirichlet(\beta^{Q_j})$
For each document x_i from Q_j:
    Draw a distribution over topics $\theta_{x_i}^{Q_j} \sim Dirichlet(\alpha^{Q_j})$
    For each word in the document $w \in x_i$:
        Draw a topic z $\sim Multinomial(\theta_{x_i}^{Q_j})$
        Draw a word $w \sim Multinomial(\phi_{x_i}^{Q_j})$

in Algorithm 3. In fact, by applying the new partitions, the vocabulary is periodically updated over time and does not become too large.

For posterior estimation we applied Online LDA proposed by Hoffman *et al.* [105]. The model is based on online variational Bayes algorithm which is based on online

stochastic optimization and known to be faster for parameter estimations with the constant running time. In this approach, a probabilistic factorization of word count matrix, which is a partition such as P_i or Q_i, represented by a matrix.

The same approach for topic inference are implemented for the two next segments. In fact, documents in new segment arrives and documents in the previous segments are fade away ,therefore, vocabulary is periodically updated in each iteration and size of process is not enlarging over time.

5.2.3 Topic Selection

In temporal topic detection, every two partitions are considered as heterogeneous sources since they were generated in different timestamps. Accordingly, the topics as the predictive variables (in our prediction model) derived from each partition are variant due to emerging information over time. We address topic evolution by ignoring topics repeated over time and selecting emerging topics in new partitions. This topic selection process allows us to select the topics which are diverse enough to represent emerging context and provides more predictive features. Topic selection is implemented in two steps. First, topic similarities are calculated and then, based on a predetermined threshold value, topics with a similarity smaller than the threshold are selected.

Similarities between topics extracted in partition P_j and Q_j are processed on a one to one level (see Figure 5.6). Two different distance measures, the Jaccard index and KL-divergence, were applied. While the Jaccard index represents information flow at word level, KL-divergence also applies word distributions. In fact, Jaccard addresses emerging words in the selection of topics and KL-divergence measures a non-symmetric relation between topics and explains how upcoming topics (K^{Q_j}) are diverse compared to the current time slice (K^{P_j}). The similarity between each topic inferred from Q_j is compared with all the inferred topics from P_j. Topic similarities for each topic k,

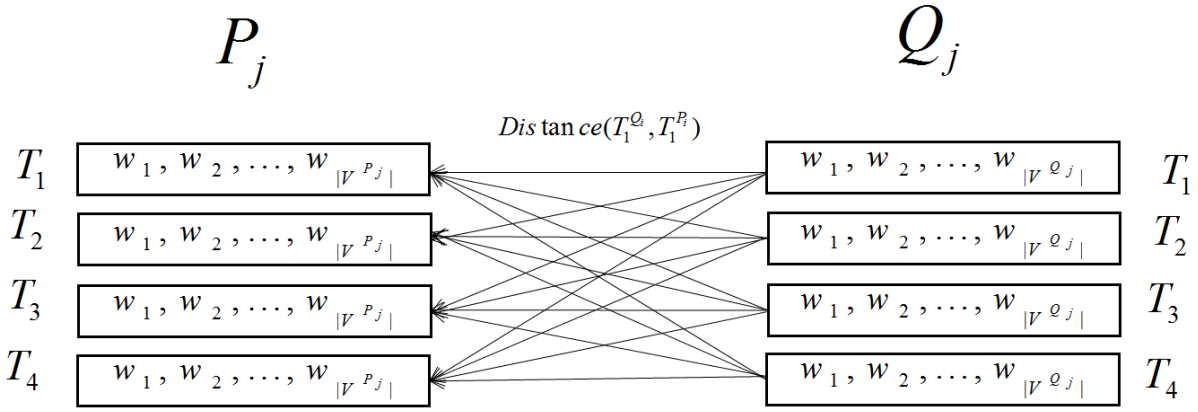


FIGURE 5.6: The general schema of topic selection with their asymmetric one to one relationships.

where $k \in K^{Q_j}$ is calculated as follows:

$$Sim(T_k^{Q_j}, T_K^{P_j}) = \sum_{h=1}^{K^{P_j}} Dist(T_k^{Q_j}, T_h^{P_j}) \tag{5.2}$$

where the distances are summed if a one to one linkage has low similarity. $Dist$ is the distance function calculated based on the Jaccard $(Dist_J)$ or KL-divergence $(Dist_{KL})$ measurements.

The next step is to rank and select divers topics. Topic diversity is measured by one of the two characteristics as follows: (i) a novel topic should have a different word distribution compared to the previous partition; or (ii) a novel topic should introduce emerging words to the dictionary which have the characteristics of not being appeared in previous segment, or having different word distributions compared to previous time slice, and bringing new emerged words in the vocabulary.

So far, for each two partitions the asymmetric one-to-one corresponding distance between topics were measured. To select the best emerging topics, a hybrid score as a linear combination of their similarity measures are calculated. The rank is given to all topics in Q by mean scores given by the distance measures.

$$Score(T_k^{Q_j}) = \frac{\sum_{h=1}^{K^{P_j}} Dist_{KL}(T_k^{Q_j}, T_h^{P_j}) + \sum_{h=1}^{K^{P_j}} Dist_J(T_k^{Q_j}, T_h^{P_j})}{2} \tag{5.3}$$

5.2.4 Document Representation

After topic inference and selection, each document (x_i) is represented by a set of novel topics. If we assume K is the total number of selected topics, each document is presented with a vector of topic distributions as follows:

$$\begin{aligned} x_i &= (T_1, T_2, ...T_K), \\ T_k &= [0, 1], 1 \leq k \leq K \end{aligned} \tag{5.4}$$

Each topic distribution is normalized with respect to the partition where the topic was inferred. As an example, if the topic was extracted from the partition P_j, then the score will be calculated considering the minimum and maximum topic distribution in P_j as follows:

$$T_k^{'P_j} = \frac{T_k^{P_j} - T_{min}^{P_j}}{T_{max}^{P_j} - T_{min}^{P_j}} \tag{5.5}$$

where $T_k^{P_j}$ refers to the topic distribution for the document x_i of partition P_j.

5.3 Results and Discussions

In this section, the experimental results are presented based on the contribution of different features: temporal topics, topics extracted from Batch LDA (LDA without time consideration), Bag-of-Word (BOW), and sentiments for predicting crime trends. In BOW model, the predictability of different smoothing windows (q) was examined. In addition, a set of experiments was conducted to study the predictability of the content compared with auxiliary features such as unemployment rates and crime rates in the past. We also present how performance is different with the availability of historical

data. For the topic model, experiments indicate that there is a need for an appropriate temporal model for detecting latent topics. It is shown how the topics are variant when inferred from the temporal model in terms of document-topic and term-topic matrices. Moreover, we also examined the predictability of topics detected by the temporal model compared with the batch model. Similar to the BOW model, the predictability over different crime types as well as different lags is presented.

5.3.1 Experimental setup

The prediction model is the proposed content-based approach which was discussed in Chapter 3. For the classifier, we use linear Support Vector Machine with its "partial_fit" function which allows online learning. For the topic identification, Online LDA proposed by Hoffman *et al.* [105] was applied. Their model uses variational Bayes (VB) for posterior inference, which has shown to be faster for large dataset analysis. We accelerate the processing time of LDA VB by using GPU-based library (BIDMach). BIDMach is a library designed to process large datasets on GPU. Table 5.1 presents how BIDMach can speed up the processing time of using LDA with the same implementation (Online LDA [105]) compared to CPU.

TABLE 5.1: Performance on GPU Vs CPU.

System	Time	Iterations	Data Throughput	Gflops	Mflops/W
BIDMach online VB (680 GPU)	40 secs	20	40 MB/s	25	250
Blei batch VB	252000 secs	20	0.1 MB/s	0.05 (est.)	0.5

The overall number of detected topics is not predefined since the proposed temporal model identifies novel topics in each iteration and adds them to the total number of final topics. Topic distributions were normalized in a range of $[0, 1]$. In the topic extraction phase, we applied differently sized partitions ranging from yearly (m = 2) to monthly (m = 12). The baseline is batch LDA with no time consideration as well as Bag-of-Word model. The number of topics for Batch model were selected between

[10,1000]. In the same manner, for BOW model, different number of features were selected and examined and the best results were reported.

The evaluation was processed by calculating the Macro-average F-measure using rolling origin approach [104] as the common method for training and evaluating the performance of the model for series observations. In this approach, the training set is the first i and it is tested on the $i+1th$ document. In the second iteration, the training set is moved one document forward (the first $i+1$), and it is tested on the $i+2th$ document. This process is continued until the test data is classified (see Figure 5.7).

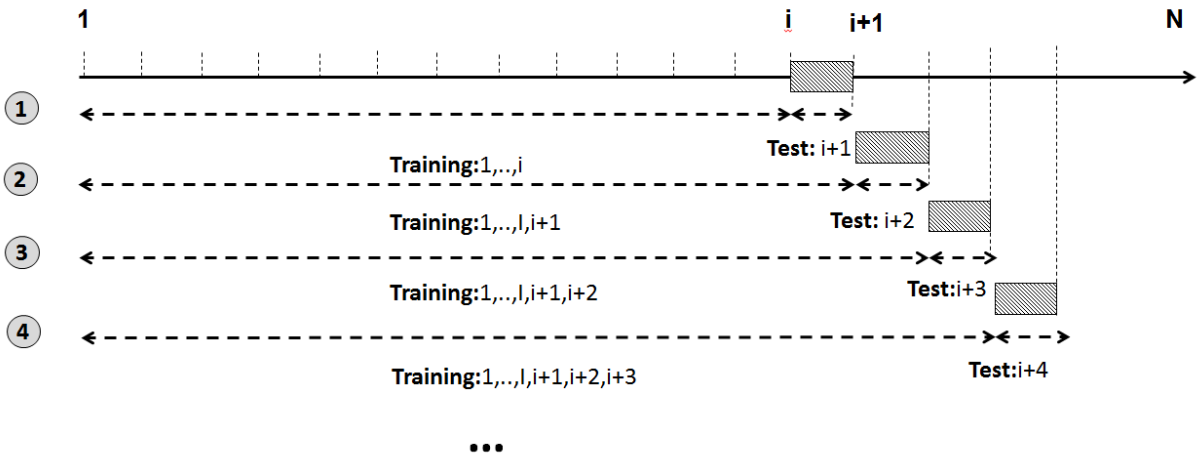


FIGURE 5.7: The division schema for rolling origin evaluation.

5.3.2 Bag-of-Word Representation

We selected bag-of-words (BOW) model as a baseline to compare it with the topic model. In BOW model, documents are texts with n-grams where $n \in \{1, 2, 3\}$. We removed stopwords and low-frequent terms. The documents were represented with a binary and tf-idf representation. The best results were achieved using $n = 1$ and binary representations. In the following subsections, we explain the results of using BOW for the targeted prediction.

5.3.2.1 Dataset Description

We collected Twitter data and crime rates from Chicago, Illinois between July 1, 2010 and November 30, 2013. Chicago has been targeted due to its importance as the third populous city in U.S as well as being among the top three cities which attracted the highest number of visitors during 2012.[1] It has been also ranked as the first in number of murders, second in robbery, and third in number of property crimes based on FBI report during 2013.[2]

Crime Data The criminal records were extracted from Chicago Data Portal [3]. This data portal is a rich resource providing all reported incidents on a daily basis which are retrieved from Chicago Police Department system. Information of all crimes which have been reported between July 2010 and November 2013 were collected. Each record contains its timestamps, exact location, and the crime type. The dates refer to the time of primary investigation, and crime type derived based on the FBI classification system. Figure 5.8 presents the crime rate time series (aggregated rates of all different crime types). The sharp spikes and troughs are coincided with some specific events and dates. However, they might be the result of missing data. A major decrease of overall crime rates is observed during the entire period of time which is started in US in 1990s [106].

Twitter Data In order to retrieve the historical tweets, a set of Twitter users was collected and historical timelines of the selected users were then retrieved and restricted to the same timeframe – between July 1, 2010 and November 30, 2013. Daily statistics of the number of posts is presented in Figure 5.9. The observed spikes in Twitter activity trend were corresponded with the important events in Chicago. The sharp spike in 2012 coincided with the presidential election in November. The high number of tweets in February 2013 is associated with Super Bowl Sunday period. The last spike is related to one day after Chicago Blackhawks won the Stanley Cup.

[1] http://en.wikipedia.org/wiki/Chicago
[2] S. Department of Justice, FBI: http://www.fbi.gov
[3] City of Chicago Data Portal: https://data.cityofchicago.org

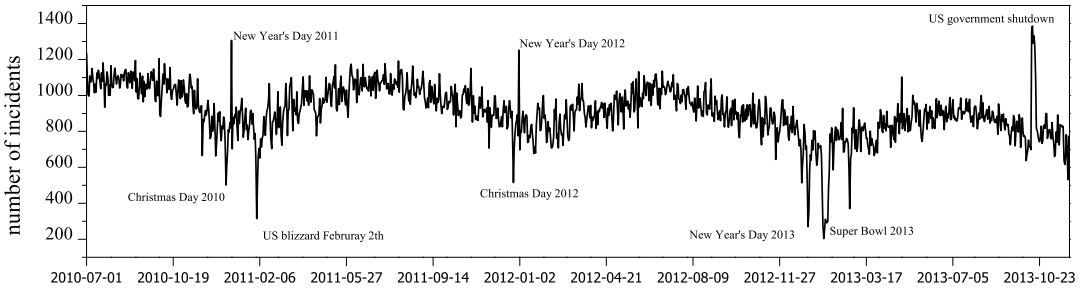


FIGURE 5.8: Daily aggregated crime rates.

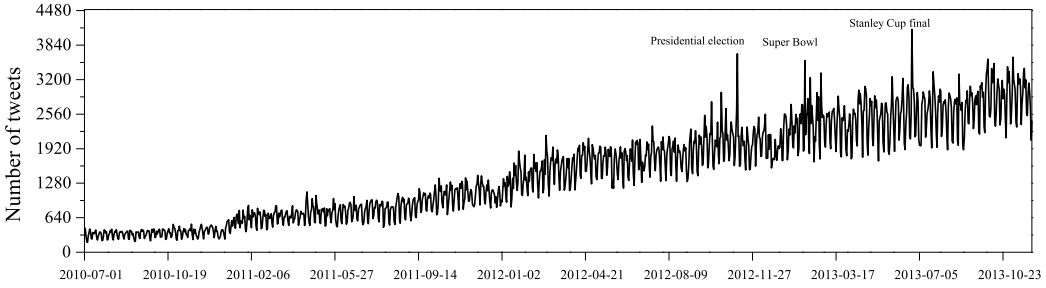


FIGURE 5.9: Daily number of tweets.

Auxiliary Resources Further, we also collected some other datasets such as unemployment rates [4] and weather conditions [5] as the auxiliary resources to investigate the expedience of their incorporation with Twitter content and to understand the contribution of content in prediction versus the other datasets.

The applied features are categorized into different groups including: content, sentiment, temporal, and auxiliary features. Table 5.2 depicts a selected list of features and the way they are employed in the classifier model.

- **Content features:** We extracted words from daily aggregated tweets. One might speculate that we must collect keywords to emphasize on offensive language implying a rough context. Nevertheless, content is a rich data which contains valuable hidden variables including activities, topic of discussions, people interests, public sentiments, which might not be carried by offensive language and should be involved in the model. In addition, some statistical features such

[4]Economic Research Federal: http://research.stlouisfed.org
[5]The Weather Channel: http://www.weather.com

as number of tweets, number of death related words, and number of swear related words are considered.

- **Sentiment features:** Sentiments captured as another set of predictive features to present general feeling of shared content on daily basis. Linguistic Inquiry and Word Count (LIWC) [92] is applied to extract daily polarity of five sentiments consist of positive, negative, anxiety, anger and sad.

  Figure 5.10 demonstrates the daily scores of the different sentiments over the observation period. The figure indicates that the overall " negative" rates have increased during the past four years compared to the other sentiments.

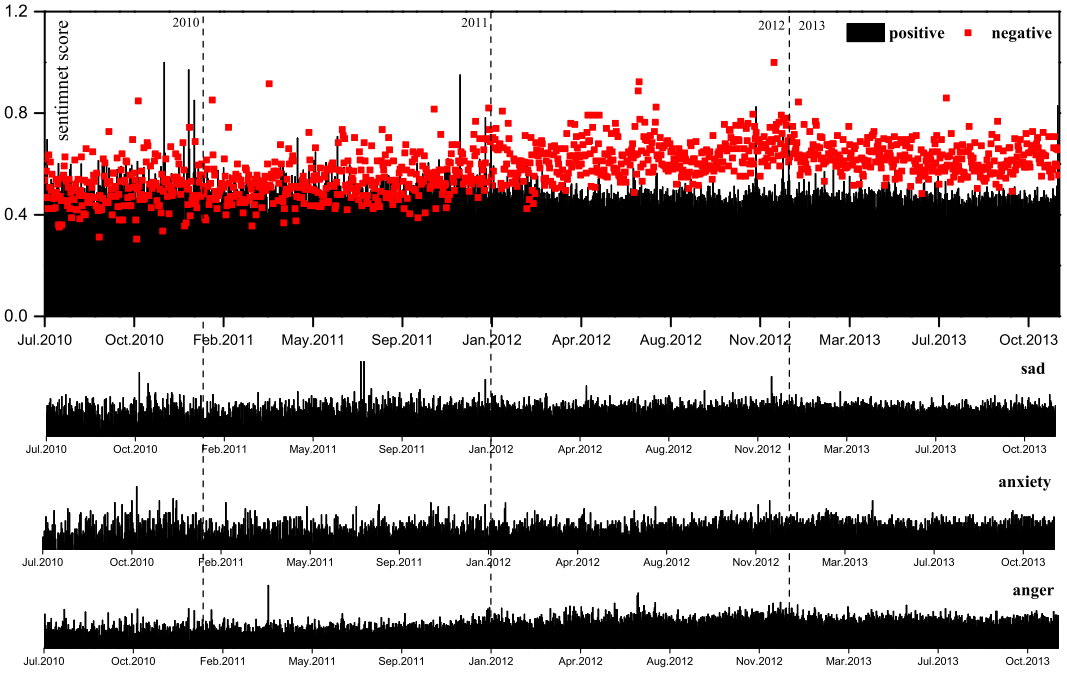


FIGURE 5.10: Sentiment scores during the observation time.

- **Temporal features:** This category consist of temporal properties of documents related to their dates. In particular, we consider whether a document is coincided with any special events or holidays.

- **Auxiliary features:** Since crime causes are various, we collected some auxiliary data such as unemployment rate as a socio-economic factor and average of climate temperature where shown to be effective in crime index prediction[107].

TABLE 5.2: List of content and auxiliary features.

Content	
Tokens	binary representation of daily aggregated tweets tweets were tokenzied stop-words and punctuations were not considered top and low frequent terminologies were removed feature selection techniques were applied: normalized IG[108] and X^2
Number of tweets	number of tweets per day
Death	[number of death related words per day/total number of words per day] $*$ 100
Swear	[number of swear related words per day/ total number of words per day] $*$ 100
Sentiment	
Positive	[number of positive words per day / total number of words per day] $*$ 100
Negative	[number of negative words per day/ total number of words per day] $*$ 100
Anxiety	[number of anxiety words per day/ total number of words per day] $*$ 100
Anger	[number of anger words per day/ total number of words per day] $*$ 100
Sad	[number of sad words per day/ total number of words per day] $*$ 100
Auxiliary	
Unemployment rate	normalized monthly unemployment rate
Weather	normalized monthly average temperature
Temporal	
Month	month bounded between 0 and 1
Labour day	3 days before and after each Labour day
Halloween	3 days before and after each Halloween
Thanksgiving	3 days before and after each thanksgiving
Christmas	3 days before and after each Christmas
New year's day	3 days before and after each New year
MLK day	3 days before and after each Martin Luther King day
Valentine's day	3 days before and after each Valentine's day
Patrick's day	3 days before and after each Saint Patrick's day
4th of July	3 days before and after each Independence day
Super Bowl	3 days before and after Super Bowl was played in February 2013
Presidential election	3 days before and after presidential election in November 6, 2012

In the following subsection, we will explain how these features are employed in our prediction model.

5.3.2.2 Smoothing Temporal Data

As discussed in Chapter 3, each temporal document (x_i) is generated using different smoothing windows (see Equation 3.3). In this part, the results of the experiment with different aggregation windows q where $q = [1, 7]$ are represented. The F-measure for each crime type is reported in Table 5.3. While the results vary based on the different crime types, daily ($q = 1$) aggregation is considered to be the best window size.

TABLE 5.3: The prediction performance based on different aggregation windows (q).

Crime type	Frequency	$q=1$	$q=2$	$q=3$	$q=4$	$q=5$	$q=6$	$q=7$
TOTAL	1,137,790	0.63	0.63	**0.64**	0.61	**0.64**	0.62	0.6
THEFT	247,617	**0.67**	0.65	0.66	0.66	0.64	0.6	0.6
BATTERY	204,041	0.78	**0.8**	0.72	0.72	0.73	0.65	0.59
NARCOTICS	124,890	**0.78**	0.74	0.71	0.62	0.63	0.64	0.59
CRIMINAL DAMAGE	120,934	**0.74**	**0.74**	0.68	0.71	0.7	0.64	0.61
BURGLARY	79,420	**0.74**	0.72	0.71	0.66	0.65	0.68	0.56
ASSAULT	65,954	0.65	**0.66**	0.61	0.63	0.61	0.62	0.59
OTHER OFFENSE	63,672	0.66	0.68	0.64	**0.69**	0.61	0.61	0.57
MOTOR VEHICLE THEFT	57,227	0.61	0.6	0.56	0.6	0.62	**0.63**	0.58
ROBBERY	4,5458	0.58	0.56	0.53	0.58	**0.6**	0.57	**0.6**
DECEPTIVE PRACTICE	40,917	0.72	**0.73**	0.69	0.64	0.65	0.63	0.56
CRIMINAL TRESPASS	28,682	**0.66**	0.65	0.64	0.64	0.6	0.61	0.61
WEAPONS VIOLATION	12,408	0.58	0.6	0.62	0.63	**0.65**	0.62	0.56
PUBLIC PEACE VIOLATION	10,661	**0.63**	**0.63**	0.63	0.61	0.6	0.58	0.58
OFFENSE INVOLVING CHILDREN	7,343	0.65	0.61	0.6	0.59	**0.72**	0.59	0.63
PROSTITUTION	7,311	**0.77**	0.75	0.71	0.57	0.7	0.6	0.61
CRIME SEXUAL ASSAULT	4,330	0.63	0.67	0.68	0.57	0.56	0.64	0.61
SEX OFFENSE	3,344	0.60	**0.63**	0.63	0.57	0.57	0.54	0.6
INTERFERENCE WITH PUBLIC OFFICER	2,982	0.57	0.6	0.55	**0.67**	0.55	0.59	0.58
GAMBLING	2,587	**0.64**	0.58	0.59	0.63	0.59	0.6	0.62
LIQUOR LAW VIOLATION	1,939	0.62	**0.68**	0.65	0.65	0.63	0.61	0.55
HOMICIDE	1,547	0.55	0.56	0.56	0.57	**0.59**	0.58	**0.59**
ARSON	1,542	**0.60**	0.57	0.57	0.56	0.55	0.53	0.58

5.3.2.3 The Impact of Historical Data

Another set of experiments was conducted to measure the impact of historical data on prediction performance. This was done to find out if the crime trend becomes more predictable as we observe more historical data or not. In this experiment, the size of test data remains unchanged (August 2013 to November 2013), and the size of training data is started from 31 days of the latest historical data (July 2013) to predict test data. In the next experiment, the size of training date is increased by sliding training window 31 more days into the past. In fact, in each experiment, the size of training data is increased by involving more documents retrospectively. The experiments are repeated until all the historical data were employed. Figure 5.11 depicts the results with the different historical training windows for all the incidents. The highest predictability is obtained when all historical data are used for the prediction model. However, the result by the seventh months is comparable to the overall performance, while adding more historical improves the performance little.

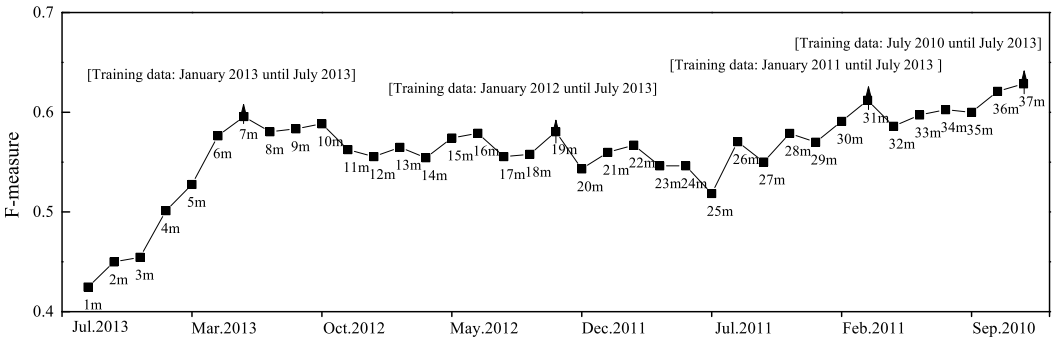


FIGURE 5.11: Test data consist of documents during August, 2013 and November, 2013. First experiment applied the training data during July 2013. For the next experiment, the training window is increased by one more month retrospectively (June 2013 and July 2013). The experiments repeated until the whole historical training data was involved. The figure indicates the F-measure for each experiment. For some of the results, the period of contributed training data presented.

5.3.3 Prediction based on Sentiment Analysis

Unlike all the previous experiments which have been conducted using content-based features, this experiment is set up to test the predictability of sentiment features. A holdout evaluation has been applied to evaluate the predictions. The experiment is conducted on all the incidents for five individual sentiment variables and one incorporated sentiment. Then we repeated the experiment by adding sentiment variable to the content. The results indicated a low predictability for sentiments. In the best case, negative sentiment, the F-measure reached up to 0.55. In fact, the sentiment analysis was not able to perform better than the content-based features in any of the cases.

5.3.4 Content Features v.s. Auxiliary Features

Although the main contribution of the paper is to study the correlation between content and crime trend, we also employ other auxiliary datasets which are widely applied in crime prediction. As discussed before, several studies have investigated the incorporation of socio-economic indexes and spatio-temporal features in crime prediction [107]. We also apply the other resources in our prediction model to understand the

contribution of the content-based features in comparison to the other predictive variables. We selected a list of non-content features (see Table 5.2), which widely applied in crime prediction. The selected features are as follows:

- **Unemployment rate:** Unemployment rates were shown to have a direct relationship with crime rates [109, 110]. These rate were leveraged as a socio-economic factor. The rates were obtained as discussed before.
- **Weather:** The normalized monthly average temperature was also employed when shown to be effective in crime index prediction [89, 107].
- **Crime rates:** Crime rates are employed as another set of features. As discussed in Chapter 2, conventional predication models employ historical crime records to predict future incidents. In our model, crime record at time t is labeled with crime records at time $t + \Delta r$, where Δr is the lag. The idea is to investigate how much a crime rate is predictive of future records.
- **Number of tweets:** The number of tweets per day is normalized between 0 and 1.
- **Day of week:** It refers to the day when a document is generated. The features are numbers from 1 to 7 where the normalize values are employed in the prediction model. We also considered month as another feature.
- **Events (Temporal):** The days before and after a set of specific events such as: Halloween, Thanksgiving, Christmas, New year's day, Martin Luther King day, Valentine's day, St Patrick's day, 4th of July, Super Bowl, and Presidential election.

We evaluated the performance of each feature as well as content-based features (unigram model). Figure 5.12 presents the performance of "day of week", " number of tweets", and "content" for predicting the increase and the decrease of crime rates. The

results indicate that content-based features significantly improve the F-measure where the other features did not provide comparable results. The rest of the features such as "unemployment rate" and "events" could not achieve high performance compared to the other auxiliary features (in the best case, F-measure = 0.4). The result indicates that, while content significantly reduced the error of prediction, auxiliary features did not contribute in prediction performance. Figure 5.12 indicates that in all presented crime types, NARCOTICS, CRIMINAL DAMAGE, BATTERY, PROSTITUTION, BURGLARY, and DECEPTIVE PRACTICE, the results of 14 different lags indicate that content of Twitter has significantly improved the prediction performance compared to "day of week" and "number of tweets". Overall, content indicates a high predictability compared to other features. The number of tweets is shown to be effective compared to day of week and crime rates.

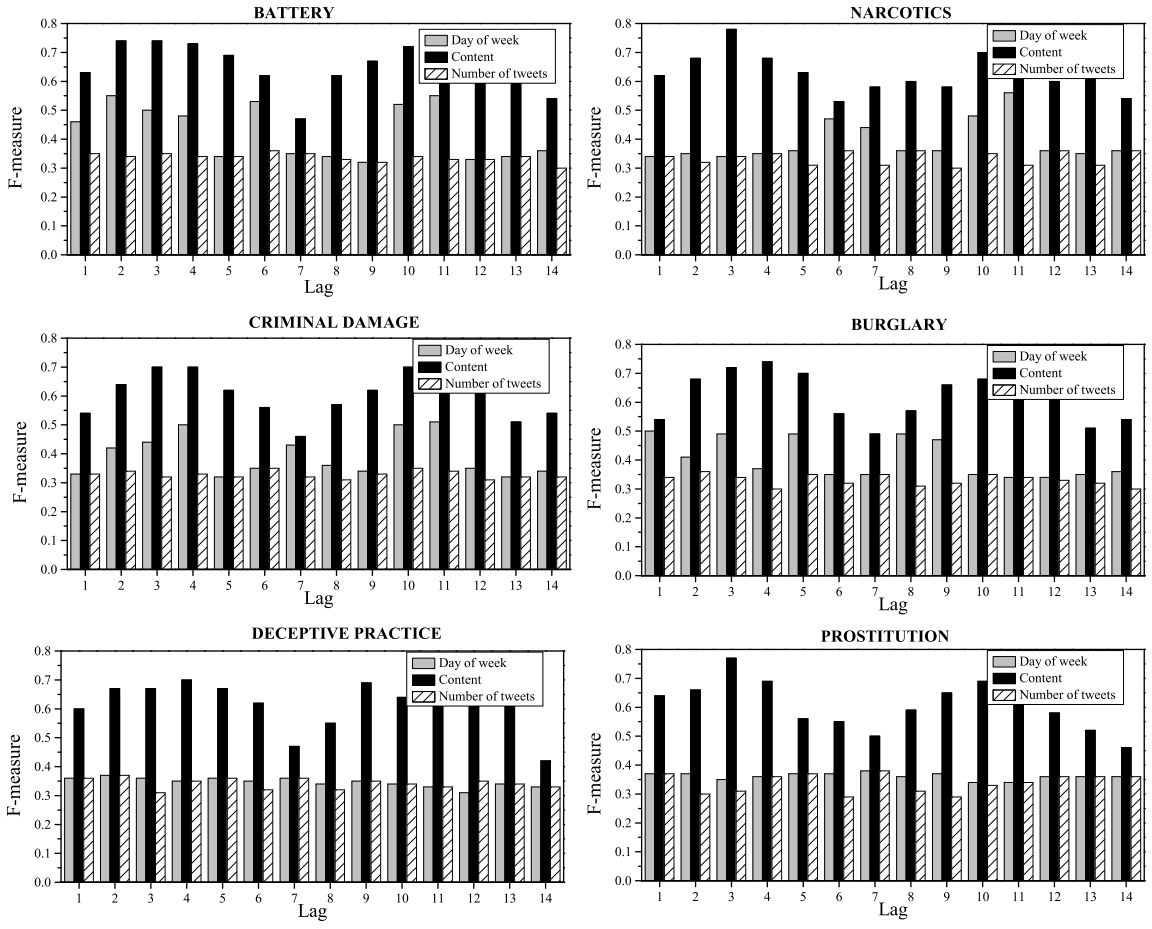


FIGURE 5.12: Performance of different features for predicting crime rate directions.

5.3.5 Prediction Performance of Temporal Topics

In this subsection, we present the experimental results of our prediction model using temporal topics. We have applied two different Twitter datasets; historical tweets in subsection 5.3.2 and sampled tweets using the proposed activity-based sampling approach (discussed in Chapter 4).

5.3.5.1 Characteristics of Temporal Topics

Identified topics from the temporal model have been compared with the baseline which is batch LDA without the time dimension. The comparison has been made in two different phases: first we compared how variant are the term distributions. Second we analyzed their differences in document-topic level.

Term-Topic Distribution: Adopting the visualization method proposed by [31], in Figure 5.13, the top 20 terms and their distributions for each individual latent topic have been visualized. The figure reveals that the topics extracted by the baseline are similar to each others as they share more similar words, while in temporal model, topics tend to have less similar terms. As shown in Figure 5.13, the vocabulary generated by the temporal model is larger compared to the baseline, therefore, more distinct topics were identified.

The second characteristic of the identified topics are topic-term distribution which has been visualized by the solid dots with different sizes. It suggests that in the temporal model, the term distribution is more variant, which means the extracted topics are more diverse compared to that of the baseline.

Document-Topic Distribution: The extracted topics show different characteristics in terms of document-topic distribution. Figure 5.14 presents the distribution of the most popular topics in each document (each day) for the batch and the temporal model. In the batch model, the extracted topics for each day has low distributions, while one

topic has shown to have high value. In most of the days, the most popular topics are topic 16 to 20 with low values. This results in poor topic identification for the whole entire period. In the temporal model, where number of partitions are between 2 to 20, the topics are fairly distributed over the documents. However, in the case of extreme partitioning, where number of partitions are 20, identified topics seem to be too general.

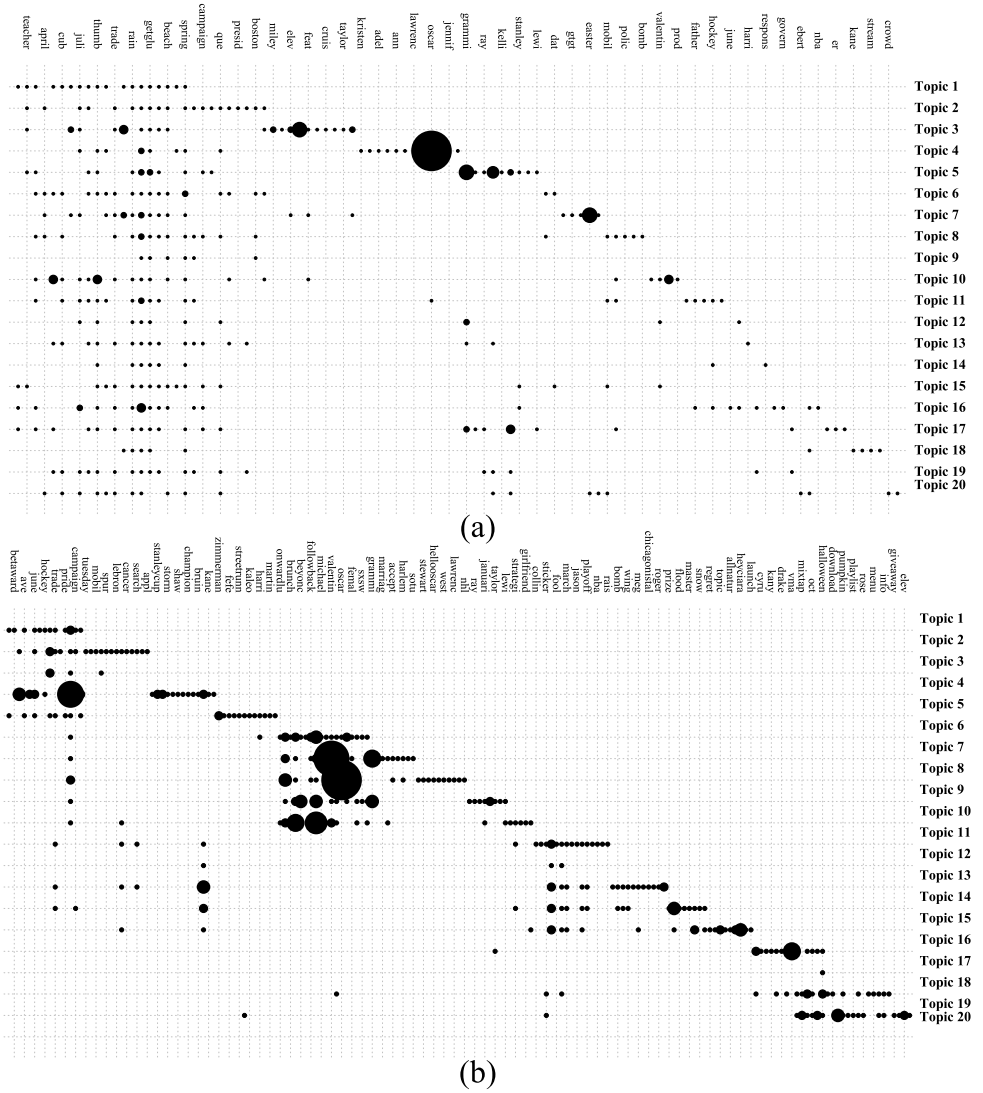


FIGURE 5.13: The most frequent terms distributions for the top 20 topics inferred by (a) baseline, and (b) temporal model.

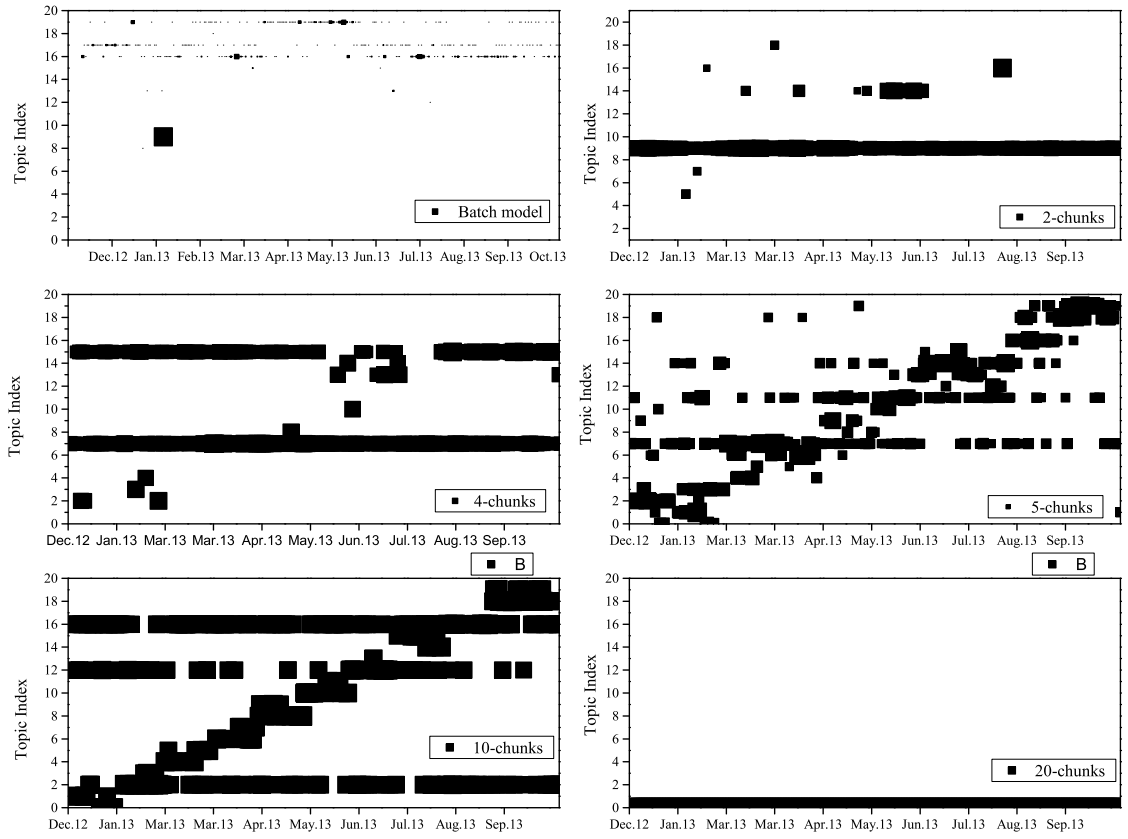


FIGURE 5.14: Topic distribution for each document based on different sizes of partition.

5.3.5.2 Temporal Topics as Features

In order to present the predictability of the temporal topic models, the experiments were expanded to 22 different crime types. For the baseline model, a predefined number of topics was observed from training corpus. In this case, any topic shift is ignored. Whereas, the temporal topic model is concerned with topic shifts and time dimension as discussed before. Table 5.4 displays the best results for each individual crime type as well as the accumulated one. It shows that the temporal model, which detects novel topics, outperformed the baseline (in 17 cases) and content (n-gram). The performance was improved by the temporal topics to 21% higher than the baseline in the best predictable crime type (Burglary). Further analysis investigated the predictability of the proposed model for different lags.

TABLE 5.4: F-measure of the best results for different crime types.

Crime type	Content	Batch model	Temporal model
ALL CRIMES	0.63	0.6	**0.76**
THEFT	0.67	0.69	**0.79**
BATTERY	0.78	0.75	**0.85**
NARCOTICS	0.78	0.7	**0.88**
CRIMINAL DAMAGE	0.74	0.67	**0.78**
BURGLARY	0.74	0.73	**0.94**
ASSAULT	0.65	**0.73**	0.7
OTHER OFFENSE	0.66	0.64	**0.7**
MOTOR VEHICLE THEFT	0.61	**0.65**	0.6
ROBBERY	0.58	0.66	**0.72**
DECEPTIVE PRACTICE	0.72	0.6	**0.73**
CRIMINAL TRESPASS	0.66	0.66	**0.67**
WEAPONS VIOLATION	0.58	**0.72**	0.67
PUBLIC PEACE VIOLATION	0.63	0.66	**0.75**
OFFENSE INVOLVING CHILDREN	0.65	0.65	**0.78**
PROSTITUTION	0.77	0.7	**0.79**
CRIME SEXUAL ASSAULT	0.63	0.72	**0.73**
SEX OFFENSE	0.60	**0.67**	0.62
INTERFERENCE WITH PUBLIC OFFICER	0.57	0.59	**0.6**
GAMBLING	0.64	**0.66**	0.54
LIQUOR LAW VIOLATION	0.62	**0.66**	**0.66**
HOMICIDE	0.55	0.63	**0.67**
ARSON	0.60	0.59	**0.73**

We also studied the predictability of the proposed model with temporal topics for different lags. A set of test scenarios were implemented to examine the predictability with different lags. Therefore, each document x_i which has been generated at time t_i is labeled with the prospective crime trends l_i. The lag does not stand for a day of week, it is a window of time in which crime rate directions are captured. Figure 5.15 illustrates the results of using temporal topics for different lags up to 7 ($\Delta r \in \{1, 7\}$). The intention is to understand the best lag between the temporal topics and crime trend. According to the results, the best performance is mostly captured when $\Delta r \in \{1, 3\}$ compared to the other lags. However, it can be variant for different crime types. Overall, the results demonstrate that the proposed prediction model with temporal topics reveals significant performance compared to other features.

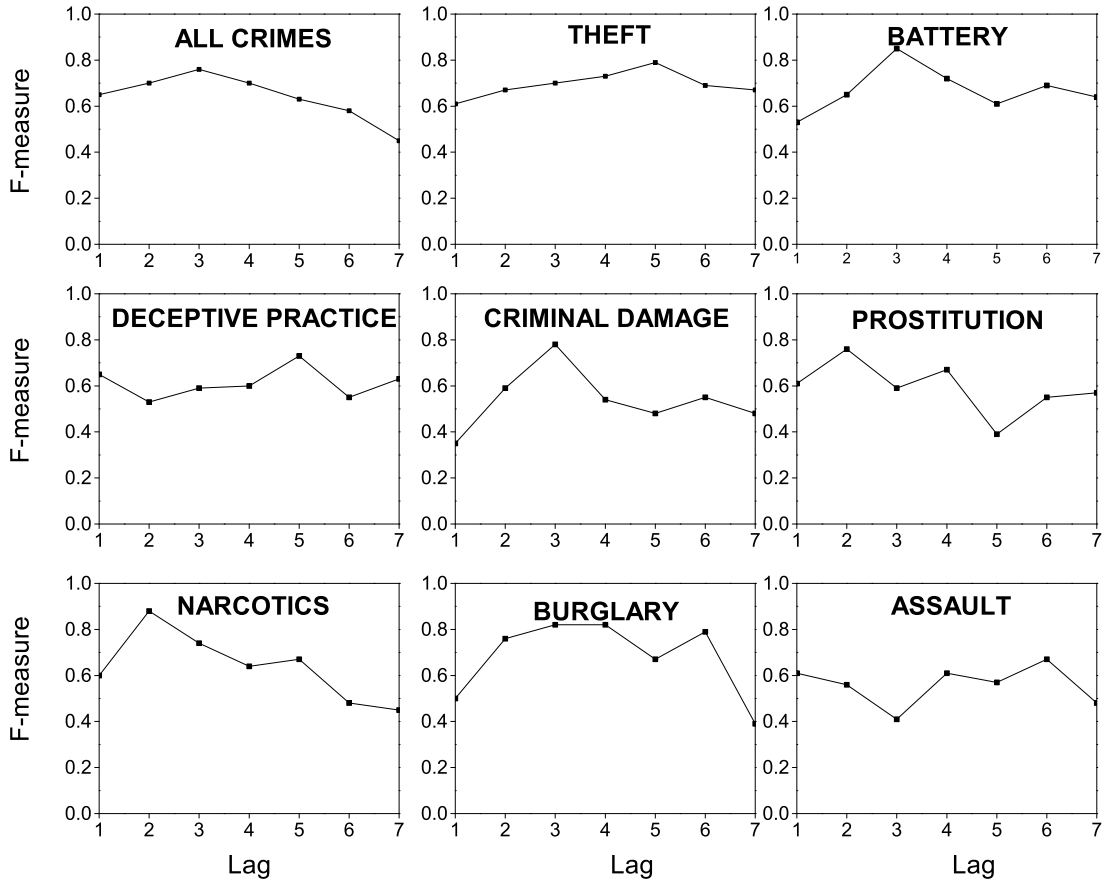
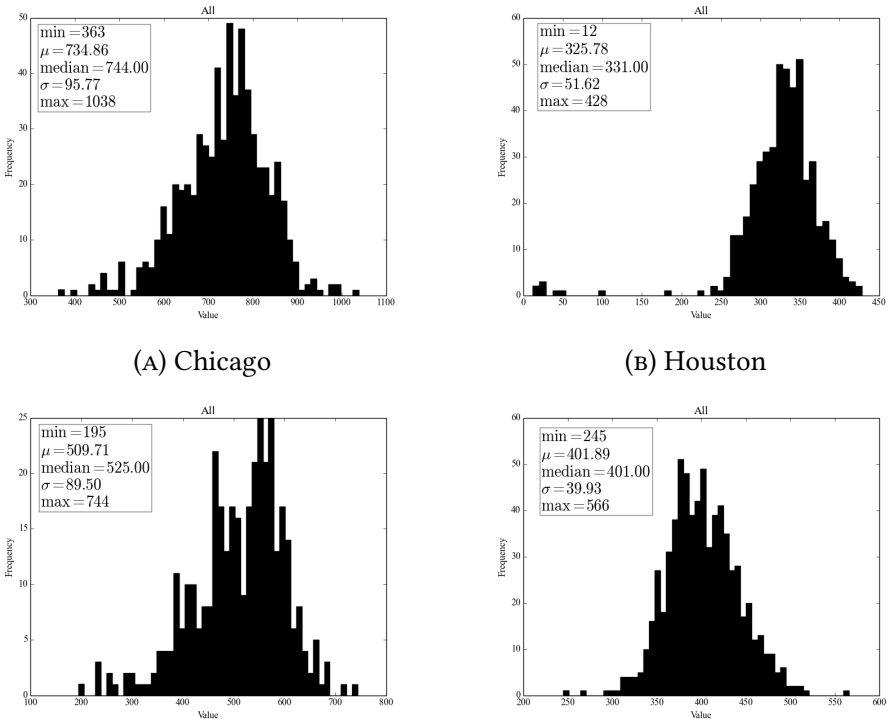


FIGURE 5.15: Holdout evaluation results for different crime types over 7 lags.

5.3.5.3 Dataset Using Activity-based Sampling

We expanded the study by applying our proposed model on different cities of United states. Crime rates were retrieved from four cities of the United States: Chicago, Philadelphia, San Francisco, and Houston. These cities were selected because they maintain a rich data portal containing crime records on daily basis. Data were collected using their data portals [6] [7] [8] [9]. For each city, different crime types were observed. Figure 5.16 presents the histogram of daily crime rates for accumulated of all incidents in different cities. In addition, tables 5.6 to 5.9 indicate crime types and their frequencies for each city.

[6] City of Chicago Data Portal: `https://data.cityofchicago.org`
[7] City of Houston eGoverment center: `http://www.houstontx.gov`
[8] SF Open Data: `https://data.sfgov.org`
[9] Phili Open Data: `https://www.opendataphilly.org`



(A) Chicago (B) Houston

(C) Philadelphia (D) San Francisco

FIGURE 5.16: Histogram of overall crime rates.

5.3.5.4 Prediction Performance

To examine the performance of predicting directions of the indexes, document x_i, which was generated at time t_i, is labeled with crime trend l_i. Documents are annotated positive or negative if the future index increases or decreases in the prospective time frame, respectively. Documents were represented with temporal topics, topics extracted from batch model, and BOW. The performance of each model is compared separately.

The predictability of temporal topics inferred with asymmetric parameters compared to trained parameters is also presented. Alpha (α) and beta (β) are the LDA parameters that affect the sparsity of document topic and topic terms matrices respectively. In the trained parameters, the parameters are learned based on one partition and are used to infer topics for the other partitions. The intention is to detect topic evolution over different partitions. As an example, if one topic was observed in partition one, the

TABLE 5.5: Crime types and frequencies (Chicago).

Crime type	Frequency
Arson	718
Narcotics	45,285
MotorVehicle	17,141
Liquor	625
Kidnapping	352
Intimidation	205
Interference	2,400
Homicide	781
Gambling	675
Deceptive	25,399
Criminal tresspass	12,346
Criminal damage	48,293
SexualAssualt	2,198
Burglary	23,492
Battery	85,916
Assualt	29,579
ChildrenOffense	3,834
OtherOffense	29,875
Prostitution	2,710
PublicViolation	4,737
Roberry	16,379
SexOffense	1,562
Stalking	259
Theft	102,522
WeaponViolation	5,681
All	462,964

TABLE 5.6: Crime types and frequencies (Houston).

Crime type	Frequency
Theft	86,793
Robbery	13,019
Rape	937
Murder	299
Burglary	27,350
Auto Theft	17,366
Aggravated Assault	12,240
ALL	158,004

model can detect the distribution of that specific topic over time. In fact, the evolution of topics are leveraged as features for the prediction model. Figure 5.17 presents the document-topic matrix for topic inference (θ) with trained parameters. In this figure, each horizontal dashed line indicates the border of each partition in the document-topic matrix. For instance, the first partition (P_1) includes documents $\{x_1, x_2, .., x_m\}$, whereas the second partition (Q_1) contains documents $\{x_{m+1}, x_{m+2}, .., x_{2m}\}$. In order to infer the distribution of topics, LDA parameters are trained based on the partition

Table 5.7: Crime types and frequencies (Philadelphia).

Crime type	Frequency
Aggravated Assault	2416
Aggravated Assault No Firearm	6218
All Other Offenses	39066
Arson	590
Burglary Non-Residential	2238
Burglary Residential	9746
Disorderly Conduct	3952
Driving Under The Influence	4695
Embezzlement	353
Forgery and Counterfeiting	264
Fraud	9178
Homicide - Criminal	330
Liquor Law Violations	582
Motor Vehicle Theft	3080
Drug Law Violations	11862
Offenses Against Family and Children	162
Other Assaults	24956
Other Sex Offenses (Not Commercialized)	1617
Prostitution and Commercialized Vice	1235
Public Drunkenness	374
Rape	887
Recovered Stolen Motor Vehicle	7365
Robbery Firearm	3375
Robbery No Firearm	4577
Theft from Vehicle	13808
Thefts	24646
Loitering	433
Criminal Mischief	16710
Weapon Violations	1527
All	196242

Table 5.8: Crime types and frequencies (San Francisco).

Crime type	Frequency
Assault	23,704
Burglary	11,085
Drug/Narcotic	8,967
Fraud	5,659
Theft	74,657
Missing Person	8,688
Non-Criminal	35,835
Other Offenses	37,877
Prostitution	752
Robbery	6,692
Secondary Codes	3,647
Stolen Property	1,906
Suspicious OCC	9,853
Vandalism	13,841
Vehicle Theft	13,887
Warrants	12,441
Weapon Laws	3,003
All	272,494

P_1 and the topics are inferred for the rest of partitions. The same approach for training parameters and inferring topics is applied for the second partition. In the second approach, asymmetric model, training parameters and inferring topics are done for each partition separately. In fact, for trained parameters, topic evolution is considered in which the distribution of each topic over time for all partitions is inferred. However, in the asymmetric model, topic distribution is only inferred in their corresponding partition.

The experimental results in tables 5.9 to 5.12 illustrate Macro-averaged F-measure obtained by temporal topics, batch LDA, and BOW for different crime types, locations, and different number of partitions ranging from 2 to 12. In these tables, "ALL" stands for the accumulation of all crime types without type consideration. Overall, the predictability is higher for Philadelphia compared to Houston, Chicago, and San Francisco. The best predictability is obtained in the case of DRUG VIOLATION (0.81) for Philadelphia, ALL (0.73) for San Francisco, BURGLARY, AUTO THEFT, RAPE (0.71) for Houston, and ALL (0.78) for Chicago. In most cases, the predictability of temporal topics is higher than BOW and batch model. Off the 25 crime types, 22 cases showed improvements when using temporal topics compared with the other two models. For Houston, this result includes all crime types and for San Francisco and Philadelphia is 14 out of 16 and 26 out of 27 respectively. In the best cases, the prediction is 18% (Chicago), 14% (Houston), 19% (San Francisco), and 33% (Philadelphia) higher than BOW and LDA. Overall, the results indicate that the temporal model is more successful in finding the most predictive topics, which can be the result of detecting more diverse and variance topics compared with Batch LDA. In addition, the results reveal that inferring topics with trained parameters (i.e topic evolution) is more effective for the predictability of the detected topics. The results present the consistency of the temporal model in delivering the best results compared to BOW and Batch LDA. Although BOW obtained the best results in a few cases such as BURGLARY in Chicago and THEFT in San Francisco, remarkable improvements in performance were obtained

by the temporal model.

In general, the results indicate that the proposed temporal model reveals a satisfactory performance for most crime types since the temporal topics are the only resource for crime prediction. However, in some types of crime such as BURGLARY in Chicago, the approach achieved the lowest result compared to the other crimes. This can be explained according to the nature of the incidents. In fact, some crimes such as BATTERY, NARCOTICS, and PROSTITUTION are mostly street incidents and might be reflected in daily social conversation while the others are more organized nature.

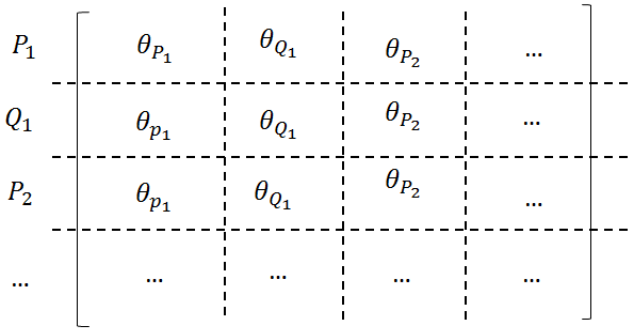


FIGURE 5.17: Temporal topic inference for trained parameters.

5.4 Conclusion

We introduced a temporal topic model to detect emerging topics as predictive variables for predicting crime trend. The proposed temporal topic model leverage temporarily nature of topics to predict the changes of crime indexes from major cities of United States, which can be extended to other areas and locations. In this approach, topics were extracted from different periods of time and were employed in a classifier model to predict crime rate changes on a daily basis. The model applies novelty and diversity in topic detection. In each period of time the best representative inferred topics are selected as features for the proposed prediction model. A comprehensive experiments

Table 5.9: Prediction performance (Chicago)

	BOW	Batch LDA	Asymmetric parameters				Trained parameters			
			2	4	8	12	2	4	8	12
ARSON	0.55	0.6	0.6	0.6	0.52	0.6	0.58	0.59	0.58	**0.64**
NARCOTICS	0.55	0.55	0.58	0.54	0.63	0.6	0.62	0.61	**0.69**	0.64
MOTOR VEHICLE	0.52	0.54	0.52	0.57	0.53	0.62	0.63	0.6	**0.65**	0.6
LIQUOR	0.56	0.57	0.6	0.58	0.53	0.57	0.61	**0.68**	0.6	0.55
KIDNAPPING	**0.63**	0.52	0.51	0.59	0.58	0.49	0.55	0.61	0.62	0.55
INTIMIDATION	0.51	0.56	0.52	0.53	0.55	0.55	0.55	**0.6**	0.59	0.54
INTERFERENCE	0.58	0.58	0.59	0.51	**0.65**	0.57	0.58	0.54	0.62	0.61
HOMICIDE	0.52	0.59	0.52	0.54	0.59	**0.65**	0.58	0.58	0.59	0.63
GAMBLING	0.59	0.55	0.56	**0.65**	0.57	0.55	0.57	0.58	0.63	0.56
DECEPTIVE PRACTICE	0.5	0.53	0.51	0.47	0.54	0.57	0.65	0.65	**0.66**	0.61
CRIMINAL TRESPASS	0.61	0.58	0.62	0.54	0.56	0.65	0.62	0.65	**0.69**	0.59
CRIMINAL DAMAGE	0.49	0.57	0.58	0.63	0.6	0.59	**0.7**	0.68	0.65	0.63
SEXUAL ASSUALT	0.64	0.54	0.54	0.6	0.52	0.57	**0.65**	0.62	0.62	0.64
BURGLARY	**0.6**	0.55	0.59	0.58	0.56	0.56	0.57	0.54	0.58	0.57
BATTERY	0.45	0.6	0.59	**0.71**	0.58	0.62	0.67	0.67	0.63	0.64
ASSUALT	0.58	0.65	0.55	0.63	0.6	0.55	0.62	0.62	**0.69**	0.58
CHILDREN OFFENSE	0.49	0.57	0.54	0.55	0.51	0.56	**0.68**	0.63	0.59	0.63
PROSTITUTION	0.55	0.59	0.65	0.51	0.59	0.59	0.64	0.67	0.61	**0.7**
PUBLIC VIOLATION	0.61	0.56	0.6	**0.62**	0.59	0.51	0.6	0.59	0.6	**0.62**
ROBERRY	0.6	0.59	0.57	0.61	0.58	0.55	0.59	**0.64**	0.63	0.63
SEX OFFENSE	0.49	**0.64**	0.52	0.52	0.56	0.56	0.6	0.58	0.59	0.57
STALKING	0.51	0.63	0.59	0.56	0.54	0.6	**0.66**	0.61	0.64	0.59
THEFT	0.6	0.57	0.54	0.61	0.55	0.59	0.6	0.61	**0.68**	0.67
WEAPON VIOLATION	0.6	0.54	0.56	0.55	0.57	0.58	0.63	0.58	0.66	**0.68**
ALL	0.55	0.6	0.55	0.57	0.61	0.57	0.73	0.71	**0.78**	0.71

Table 5.10: Prediction performance (Houston)

	BOW	Batch LDA	Asymmetric parameters				Trained parameters			
			2	4	8	12	2	4	8	12
THEFT	0.59	0.55	0.52	0.52	0.54	0.53	0.6	0.59	0.58	**0.63**
ROBBERY	0.59	0.57	0.52	0.53	0.6	0.54	0.59	0.52	0.68	**0.69**
RAPE	0.57	0.53	0.61	0.58	0.53	0.54	0.58	**0.71**	0.67	0.6
MURDER	0.52	0.59	0.57	0.53	0.53	0.61	0.49	0.46	0.5	**0.68**
BURGLARY	0.6	0.57	0.57	0.51	0.61	0.55	0.63	**0.71**	0.69	0.6
AUTO THEFT	0.49	0.58	0.53	0.61	0.64	0.61	0.64	0.53	**0.71**	0.59
ASSAULT	0.54	0.61	0.55	0.51	0.61	0.62	**0.68**	0.65	0.59	0.58
ALL	0.57	0.48	0.63	0.47	0.48	0.52	**0.64**	0.56	0.53	0.57

Table 5.11: Prediction performance (San Francisco)

	BOW	Batch LDA	Asymmetric parameters 2	4	8	12	Trained parameters 2	4	8	12
ASSAULT	0.52	0.5	0.54	0.48	**0.7**	0.57	0.64	0.64	0.6	0.63
BURGLARY	0.58	0.57	0.51	0.62	0.53	**0.65**	0.64	0.61	0.5	0.64
DRUG/NARCOTIC	0.58	0.59	0.54	0.54	0.54	0.5	0.58	0.53	0.6	**0.7**
FRAUD	0.61	0.53	0.58	0.45	0.57	0.52	0.55	**0.63**	0.61	0.56
LARCENY/THEFT	0.57	0.54	0.54	0.56	0.65	0.55	0.65	**0.69**	0.65	0.68
MISSING PERSON	0.54	0.63	0.56	0.56	0.61	0.54	0.54	0.54	0.59	**0.64**
PROSTITUTION	0.55	0.61	0.63	0.56	0.63	0.58	**0.68**	0.5	0.58	0.57
ROBBERY	0.63	0.51	0.5	0.54	0.62	**0.64**	0.6	0.59	0.63	0.65
SECONDARY CODES	0.64	0.59	0.58	0.54	0.51	0.6	0.53	0.54	**0.65**	0.64
STOLEN PROPERTY	**0.62**	0.51	0.55	0.55	0.52	0.59	0.59	0.54	0.51	0.57
SUSPICIOUS OCC	0.59	0.57	0.57	0.59	0.59	0.57	**0.65**	0.6	0.61	0.57
VANDALISM	0.54	0.57	0.6	0.6	0.57	0.54	0.56	0.6	**0.62**	0.61
VEHICLE THEFT	**0.65**	0.51	0.54	0.55	0.6	0.54	0.61	0.57	0.58	0.54
WARRANTS	0.54	0.56	**0.67**	0.63	0.48	0.63	0.63	0.61	0.59	0.6
WEAPON LAWS	0.63	0.55	0.62	0.55	0.54	0.63	0.63	0.64	0.55	**0.66**
All	0.52	0.54	0.54	0.6	0.53	0.62	0.61	0.68	**0.73**	0.61

Table 5.12: Prediction performance (Philadelphia)

	BOW	Batch LDA	Asymmetric parameters 2	4	8	12	Trained parameters 2	4	8	12
ASSUALT	0.57	0.49	0.58	0.58	0.62	0.57	0.63	0.69	**0.7**	0.6
ARSON	0.52	0.49	0.54	0.56	0.51	0.62	0.53	0.61	0.56	**0.66**
BURGLARY Non-Residential	0.61	0.5	0.61	0.53	0.5	0.54	0.61	0.66	0.63	**0.67**
BURGLARY Residential	0.51	0.47	0.58	0.49	0.52	0.55	0.66	0.64	**0.76**	0.61
DISORDERLY CONDUCT	0.54	0.48	0.63	0.49	0.51	0.52	0.66	0.58	0.61	**0.67**
DRIVING UNDER THE INFLUENCE	0.41	0.46	0.5	0.5	0.56	0.5	0.72	0.75	0.66	**0.76**
EMBEZZLEMENT	0.66	0.49	0.52	0.57	0.54	0.53	0.5	0.52	**0.74**	0.66
FORGERY AND COUNTERFEITING	0.5	0.53	0.52	0.54	0.55	0.59	0.52	0.56	0.61	**0.63**
FRAUD	0.68	0.54	0.55	0.54	0.48	0.51	0.7	**0.73**	0.71	0.72
HOMICIDE	0.62	0.48	0.63	0.62	0.54	0.53	**0.65**	0.63	0.59	0.63
LIQUOR LAW VIOLATIONS	0.52	0.46	0.57	0.58	0.52	0.56	0.54	0.58	**0.68**	**0.68**
MOTO VEHICLE THEFT	0.47	0.48	0.56	0.69	0.55	0.52	0.6	0.61	0.69	**0.72**
DRUG LAW VIOLATIONS	0.44	0.48	0.49	0.59	0.52	0.49	0.68	0.78	0.7	**0.81**
OFFENSES AGAINST FAMILY AND CHILDREN	0.43	0.47	0.49	0.61	0.55	0.48	**0.64**	0.54	0.56	0.62
OTHER SEX OFFENSES	**0.66**	0.55	0.56	0.49	0.48	0.57	0.61	**0.66**	0.56	0.62
PROSTITUTION	0.58	0.57	0.5	0.58	0.56	0.49	**0.78**	0.65	0.68	0.73
PUBLIC DRUNKENNESS	0.56	0.45	0.57	0.54	0.54	0.56	0.56	0.65	**0.67**	0.63
RAPE	0.6	0.48	0.59	0.55	0.58	0.57	0.62	**0.68**	**0.68**	0.61
STOLEN VEHAICLE	0.48	0.5	0.56	0.59	0.61	0.52	0.58	0.62	0.61	**0.63**
ROBBERY FIREARM	0.59	0.5	0.54	0.57	0.55	0.57	0.58	**0.66**	0.65	0.62
ROBBERY NO FIREARM	0.61	0.48	0.57	0.62	0.5	0.52	0.66	**0.67**	0.54	0.62
THEFT FROM VEHICLE	0.63	0.65	0.62	0.59	0.64	0.63	0.63	0.64	0.52	**0.71**
THEFTS	0.64	0.59	0.62	0.64	0.61	0.59	**0.67**	0.61	0.64	0.62
LOITERING	0.57	0.58	0.58	0.61	0.6	0.62	0.65	0.66	**0.73**	0.64
CRIMINAL MISCHIEF	0.67	0.65	0.63	0.66	0.64	0.67	0.69	0.67	**0.71**	0.63
WEAPON VIOLATIONS	0.58	0.59	0.58	0.59	0.61	0.62	0.57	**0.66**	0.63	**0.66**
ALL	0.58	0.62	0.61	0.63	0.59	0.58	0.61	0.61	0.66	**0.7**

was conducted to evaluate the performance of temporal topics compared to static topics and BOW. The conclusion of the results were presented in Table 5.13. The table shows in how many cases (crime types) each model obtained the best results compared to others. Although, BOW and batch LDA performed similarly in prediction, The performance of temporal model in 71 and 75 crime types surpassed BOW and batch LDA, respectively. The results clearly indicate that the temporal topic detection is capable of finding the predictive set of topics in each partition which greatly affect the quality of the prediction model compared to the baseline.

TABLE 5.13: BOW vs LDA vs Temporal model (overall results).

	BOW	Batch LDA	Temporal model
BOW		42-33	5-71
Batch LDA	33-42		1-75
Temporal Model	71-5	75-1	

Chapter 6

Conclusion and Future Work

This thesis has presented the idea of applying social media data, Twitter data, for trend prediction, in particular crime trend prediction. Many twitter-driven prediction models were developed, however, crime prediction based on Twitter data is less explored. This study presented different models for crime trend prediction based on mining tweets posted from a relevant geographic area. The proposed prediction method does not need any manually generated training data and it annotates its required data. In fact, the model generates its own training data by employing the knowledge inferred from targeted signals (in this case, crime indexes) and then labels are assigned to the input data. Using the proposed models, predictive features were extracted such as sentiment, BOW, and topics. In order to infer topics, a temporal topic inference model was presented which employs the changes of terms in the vocabulary to infer emerging topics. For Twitter data collection, an activity-based sampling approach was developed to avoid activity gaps over time. In the conducted experiments, temporal topics achieved the highest predictability compared to content and sentiment. Overall, the study supported the importance of considering Twitter content as an extra data resource without suffering from unlabeled data.

6.1 Conclusion

In chapter 1, we explained how we defined the targeted problem and how we approached it from different prospectives. Chapter 2 discussed existing studies in twitter-driven models, Twitter sampling, temporal topic detection, as well as crime prediction. Chapter 3 introduced two prediction models (content-based and user-centric) as well as how data were annotated for trend prediction. The frameworks of both models were presented in Figure 3.2. The models are based on extracting predictive features from user-generated content. In the user-centric model, the signals or features are sentiments which are driven from timelines of a set of selected users, to extract meaningful patterns. However, in the content-based approach, features are driven from the content of all individuals. Both of the aforementioned approaches are considered to be temporal models, which suffer from the challenge of retrieving tweets over time. In a temporal model where content is tracked to detect a set of patterns, the availability of tweets over time significantly effects the models performance. In fact, temporal models suffer from activity gaps or missing data. Therefore, In Chapter 4, we introduced a sampling approach to detect more credible users while mitigating the effect of missing content. The data gathered using the proposed sampling was evaluated on the proposed prediction models (discussed in Chapter 3). The data was evaluated by three criteria: the number of available data (the number of tweets and users) over time, activity of users (whether they are present or absent over the consideration period of time), and prediction performance. The results indicate that the proposed sampling approach has better coverage compared to the random sampling. In terms of sparsity of users activities, active users had more contribution in the past compared to the random users. Overall, the activity-based approach identifies users who are more historically active, whereas in the random sampling high activity gaps are observed. Moreover, the prediction performance of collected data was studied and the findings indicated that the content of active users achieved significantly higher performance in crime trend prediction.

In chapter 4, we proposed a temporal topic model to detect emerging topics as predictive features for crime trends prediction. In fact, after the evaluation of the proposed sampling approach in Chapter 3, we collected content of active users. The predictive signals were extracted from their content which includes, bag-of-words, sentiments, and topics. We also evaluated the contribution of auxiliary features such as weather, time of the year, unemployment rates. The results indicate that discussed topics among users achieved highest performances compared to other features. However, our problem has a sequential order and extracting meaningful patterns involves temporal analysis. We presented a temporal topic detection model to infer temporal topics. The model builds a dynamic vocabulary to detect emerging topics. Topics are compared over time to have diversity and novelty in each time consideration. The experiments have revealed that temporal topic detection outperforms static topic modeling, BOW, and other features. In addition, the characteristics of the emerging topics compared to static topics indicate that topics are more diverse when they are inferred using the proposed temporal model.

6.2 Challenges and Future Works

The aim of this thesis was mainly to propose effective approaches, techniques and algorithms for the challenges of detecting predictive features, tackling missing data, and handling temporality nature of content for the problem of crime trend prediction. Working on the mentioned challenges, we found out some interesting problems that can be addressed in the future. The future works are as follows:

6.2.1 Semantic Analysis of Twitter Sampling

In this study, content comes as a temporal stream and the problem of missing data is inevitable. Historical content is the result of retrieving the timelines of a set of users

in the past, in which the number of data available over time is inconstant. The activity gaps for both content-centric and user-centric models can mislead and degrade models performances. The proposed activity-based sampling approach can detect users with more active days in their histories. This research has shown the importance of a target-oriented data sampling for prediction models. In addition to the timeline properties and the credibility, we would like to further investigate the quality of the content in terms of discussion topics and sentiments to semantically analyze textual content and their differences in content level. The topics can be evaluated in terms frequencies level rather than only focusing on document-topic distribution. In addition, future studies can analyze syntactic structure to grammatically analyze the textual data. Future work could also address the effectiveness of the proposed sampling approach for other temporal prediction models.

6.2.2 Time-discrete Topic Detection Model

Our proposed temporal topic detection model is time discrete. Topics are inferred over discrete time slices (partitions) and topics extracted in each partition are compared to the previous partition (sequentially ordered). In fact, co-occurrences of terms in one partition is compared to the occurrences of the same words in the previous partition. Also, in time-discrete topic detection, the overall number of detected topics is not fixed since topics have birth, death, and rebirth. However, the approach of partitioning documents, conditioned on time, depends on the selection of a proper partition size. If the size of the partition is large, detected topics may suffer from high frequent terms. For a very small partition size, this leads to more computation time and may not effectively capture a significant topic due to the lack of information. Overall, the problem is having a fixed size of timestamps while topics are evolved at different speeds. One term stays popular for a very long time, while others are well-known for shorter periods. A time-continuous extension of topic inference gives the flexibility to detect temporal changes of topics disregarding the limitation of fixed term distributions in a specific

period of time. However, developing such a system can be challenging due to tuning Dirichlet parameters for topic inference. In fact, a time-continuous topic detection is similar to online learning, in which some important issues need to be significantly considered. The challenges are: when to update parameters, which data to keep for extracting topics, and when to re-train the LDA model to infer the topics.

6.2.3 Deep Structured Learning

Since performance of machine learning algorithms highly depends on the choice of data representation [111], data engineering and feature selection play a significant role in this regard. On the other hand, recent acquisition of deep learning in many NLP tasks, such as entity recognition [111], sentiment analysis [112], and POS tagging [113], indicates its effectiveness in presenting semantic representation of text documents without data engineering. Although the predictability of some variables derived from Twitter content was successfully proven in this study, further analysis by extracting other informative signals may be undertaken. We would like to semantically analyze textual content for better understanding of the relationship between features.

6.2.4 Applications – other Socio-economic Indexes

In this thesis, the proposed models were effective in crime trend prediction, however, it is interesting to investigate the effectiveness of the models for other socio-economic indexes. The proposed activity-based sampling approach can be applied for the collection of credible users for different applications such as experts in stock market. In addition, prediction models, the content-based and the user-centric, can be applied on trend prediction of other socio-economic indexes, such as unemployment rates and stock market.

Appendix A

List of Symbols

D	a set of temporal documents
d_i	a set of posts shared at t(i)
f	a feature in the global vocabulary
K	total number of topics, $1 < k < K$
K^P	total number of topics for partition P
K^Q	total number of topics for partition Q
l_i	label of a document at t(i)
M	total number of users
m	size of a partition
N	total number of documents
p_i	a post tweeted at time $t(i)$
P	documents grouped in partition t
Q	documents grouped in partition $t + m$
q	aggregation window
s_u	sentiment score belongs to user (u)
$t(i)$	timestamp of document i
T	a topic distribution
T^{P_J}	a topic distribution detected from partition P_j

T^{Q_J}	a topic distribution detected from partition Q_j
u_i	ith user out of M
V	global vocabulary
w	a term in a vocabulary
x_i	feature vector of the i-th document
X	a set of documents
$X^{\langle c \rangle}$	document term matrix of size $N * \|V\|$ sparse matrix
$X^{\langle u \rangle}$	document sentiment matrix of size $N * M$ sparse matrix
y_i	crime rate at time $t(i)$
Z_x	a topic in document x
α	hyper parameter for per document topic proportion
β	hyper parameter for per topic word distribution
θ_x	topic distribution for document
ϕ	word distribution for topic
Δr	lag between a document and a target trend

Appendix B

The Most Probable Terms for Topics

Table B.1: The most probable terms for topics extracted from batch LDA. The threshold of distribution more than 0.001 has been applied.

Topic1	hall	teacher	heat	whackstar	cub	mixtap	juli	snow	thumb	march	coast	rain	hawk	getglu	footbal	beach	lolla	spring		
Topic2	teacher	april	juli	snow	trade	rain	hawk	getglu	footbal	beach	spring	fest	campaign	sticker	que	appli	presid	rais	boston	justin
Topic3	teacher	mixtap	juli	trade	coast	hawk	getglu	footbal	beach	justin	miley	cyru	elev	vma	feat	gaga	cruis	nsync	taylor	gunplay
Topic4	juli	thumb	march	hawk	getglu	lolla	spring	que	kristen	hathaway	adel	stewart	ann	speech	lawrenc	pro	oscar	ben	jennif	hellooscar
Topic5	teacher	heat	snow	thumb	rain	hawk	getglu	footbal	spring	campaign	sticker	grammi	raven	ray	beyonc	kelli	superbowl	stanley	boppin	lewi
Topic6	heat	april	whackstar	cub	snow	thumb	march	trade	rain	hawk	getglu	footbal	beach	spring	que	appli	boston	justin	marathon	dat
Topic7	april	mixtap	juli	march	trade	coast	rain	hawk	getglu	footbal	beach	spring	elev	feat	gunplay	church	gtgt	egg	easter	cancer
Topic8	heat	april	cub	trade	rain	hawk	getglu	footbal	beach	spring	fest	campaign	que	boston	marathon	mobil	soundcloud	polic	race	bomb
Topic9	hawk	getglu	beach	spring	fest	boston														
Topic10	heat	whackstar	cub	juli	snow	thumb	trade	rain	hawk	getglu	beach	spring	appli	justin	feat	soundcloud	htt	valentin	parad	prod
Topic11	heat	cub	juli	snow	thumb	march	rain	hawk	getglu	footbal	spring	fest	oscar	mobil	soundcloud	unlock	father	ahead	hockey	followback
Topic12	juli	thumb	rain	hawk	getglu	spring	que	grammi	valentin	obama										
Topic13	whackstar	cub	snow	thumb	trade	rain	hawk	getglu	footbal	spring	fest	campaign	appli	rais	grammi	beyonc	harri			
Topic14	thumb	rain	hawk	getglu	footbal	spring	hockey	respons												
Topic15	hall	teacher	thumb	march	trade	rain	hawk	getglu	footbal	beach	lolla	spring	campaign	que	stanley	dat	mobil	valentin		
Topic16	hall	heat	juli	thumb	trade	rain	hawk	getglu	footbal	beach	fest	campaign	stanley	father	hockey	june	obama	octob	halloween	govern
Topic17	hall	heat	whackstar	snow	thumb	trade	rain	hawk	getglu	footbal	spring	campaign	que	grammi	soundcloud	superbowl	ray	raven	halftim	er
Topic18	coast	rain	hawk	getglu	spring	stanleycup	kane	anniversari	stream	becauseitsthecup										
Topic19	whackstar	cub	juli	snow	thumb	trade	rain	hawk	getglu	footbal	spring	fest	que	rais	ray	beyonc	superbowl	octob	wing	
Topic20	april	cub	juli	thumb	march	trade	rain	hawk	beach	spring	que	beyonc	easter	cancer	mobil	ebert	stanleycup	superbowl	crowd	downtown

Table B.2: The most probable terms for topics extracted from temporal model with two partitions. The threshold of distribution more than 0.001 has been applied.

Topic1	heat	april	whackstar	winter	opportun	snow	cub	appli	thumb	march	respons	februari	harri	htt	hawk	oscar	pope	intern	girlfriend	jennif
Topic2	april	whackstar	snow	appli	thumb	march	hawk	oscar	jennif	grammi	carpet	hathaway	adel	stewart	ann	hellooscar	seth	graduat	lawrenc	kristen
Topic3	april	snow	appli	thumb	march	hawk	bostonmarathon	pray	runner	regret	version	marathon	pro	topic	valentin	transit	explos	boston	prayer	bomb
Topic4	april	snow	cub	thumb	hawk	marathon	boston	bomb	manhunt	aliv	playoff	polic	terror	rain	custodi	cnn	rsd	suspect	area	boat
Topic5	april	whackstar	opportun	snow	appli	thumb	march	hawk	intern	marathon	valentin	boston	rain	femal	plu	seri	wing	schedul	easter	hockey
Topic6	heat	april	cub	thumb	march	hawk	grammi	graduat	valentin	rain	seri	easter	church	draft	heyciara	egg	jesu	brunch		
Topic7	heat	april	whackstar	winter	opportun	snow	thumb	march	hawk	oscar	pope	intern	grammi	graduat	rain	wing	jack	suit	taylor	nba
Topic8	heat	whackstar	snow	thumb	march	htt	hawk	valentin	hockey	onwardlu	presid	click	inaug	inaugur	obama	rambler	rockboybam	gffchi	patrick	nhl
Topic9	heat	april	whackstar	winter	opportun	snow	cub	appli	thumb	march	hawk	intern	graduat	valentin	rain	seri	wing	dat	beyonc	angel
Topic10	april	whackstar	winter	snow	cub	thumb	march	hawk	boston	rain	area	wing	hockey	obama	entertain	followback	june	seabrook	jordan	meg
Topic11	rain	justin	mixtap	award	juli	djpierr	octob	winner	prod	getglu	lolla	halloween								
Topic12	lolla																			
Topic13	boston	rain	hockey	miley	father	award	juli	octob	getglu	halloween	parad	stanley	teacher	download	govern	shutdown	vma	pumpkin	sticker	oct
Topic14	hockey	nba	june	jame	award	juli	octob	mobil	parad	stanley	spur	firework	lebron							
Topic15	juli	getglu	lolla	parad	stanley															
Topic16	rain	obama	award	juli	octob	getglu	halloween	govern	campaign											
Topic17	rain	dat	meg	justin	august	award	juli	djpierr	octob	getglu	halloween	download	firework	campaign	moon	donat	smwchicago	pride	boppin	wavefront
Topic18	click	justin	miley	august	mixtap	award	octob	prod	getglu	que	septemb	vma	elev	feat	coast	sharknado	ave	mtvhottest	scienc	gunplay
Topic19	rain	june	award	juli	getglu	bbq	firework	scienc	independ	fourth										
Topic20	harri	rain	dat	june	justin	august	award	juli	djpierr	getglu	creativ	lolla	mobil	stanley	que	septemb	campaign	research	extra	camp

Table B.3: The most probable terms for topics extracted from temporal model with four partitions. The threshold of distribution more than 0.001 has been applied.

Topic1	justin	betaward	ave	june	diehard	hockey	lolla	trade	pride	hawk	campaign	parad								
Topic2	stanley	june	lolla	trade	hall	hawk	campaign	tuesday	data	mobil	lollapalooza	spur	schedul	lebron	appli	cancer	produc	search	test	appl
Topic3	lolla	hawk	lollapalooza																	
Topic4	stanley	father	june	hockey	hawk	campaign	parad	letsgohawk	stanleycup	becauseitsthecup	storm	crawford	shaw	toew	champion	firework	bruin	boston	kane	royal
Topic5	justin	ave	june	lolla	trade	pride	hawk	parad	zimmerman	trayvon	fefe	cori	streetrunn	feat	kaleo	guilti	harri	georg	martin	phish
Topic6	hawk	harri	onwardlu	blackhawk	brunch	whackstar	beyonc	goodfoodchi	followback	thumb	michael	februari	valentin	obama	oscar	pope	femal	feb	sxsw	presid
Topic7	hawk	blackhawk	whackstar	followback	thumb	februari	valentin	obama	oscar	pope	femal	grammi	vday	marriag	kiss	accept	valentinesday	harlem	senat	sotu
Topic8	hawk	blackhawk	whackstar	followback	thumb	februari	oscar	feb	accept	harlem	hathaway	stewart	ann	hellooscar	tie	west	goodmorn	lawrenc	jennif	nhl
Topic9	hawk	blackhawk	whackstar	beyonc	thumb	valentin	obama	pope	feb	sxsw	presid	grammi	raven	ray	gffchi	januari	superbowl	taylor	goldenglob	lewi
Topic10	hawk	appli	onwardlu	blackhawk	whackstar	followback	thumb	februari	valentin	obama	feb	presid	kiss	januari	lewi	freshcont	strategi	jan	girlfriend	chichat
Topic11	trade	appli	search	boston	strategi	collin	wrigley	sticker	april	fool	rain	march	unlock	jason	opportun	playoff	donat	nba	sponsor	rais
Topic12	boston	april	rain																	
Topic13	trade	appli	search	boston	april	rain	march	opportun	playoff	sox	bomb	marathon	wing	suspect	meg	ebert	chicagonistal	polic	roger	easter
Topic14	trade	campaign	boston	strategi	april	rain	march	opportun	playoff	bomb	marathon	wing	prize	smssummit	flood	gno	master	pro	snow	recruit
Topic15	appli	boston	chichat	april	rain	march	opportun	meg	smssummit	pro	recruit	regret	earth	topic	twittertalk	allnatur	transit	heyciara	select	launch
Topic16	taylor	miley	cyru	nsync	kany	que	drake	vma	getglu	oct	award	halloween								
Topic17	halloween																			
Topic18	obama	sticker	rain	miley	que	sept	govern	mixtap	getglu	oct	halloween	teacher	download	pumpkin	nfl	soundcloud	menu	shutdown	info	cub
Topic19	feat	sticker	govern	mixtap	getglu	oct	award	halloween	coast	pumpkin	brown	playlist	nfl	rose	shutdown	info	giveaway	smwchicago	elev	gunplay
Topic20	fool	march	rain																	

Table B.4: The most probable terms for topics extracted from temporal model with five partitions. The threshold of distribution more than 0.001 has been applied.

Topic1	mobil	soundcloud	menu	chicagonistal	relationship	shutdown	download	marathon	dat	rain	govern	smwchicago	sticker	cub	nfl	oct	septemb	award	info	sept
Topic2	angel	ebert																		
Topic3	rain	septemb	juli																	
Topic4	nate	snow	march	twittertalk																
Topic5	donat	issu	trade	angel																
Topic6	letsgohawk	stanley	stanleycup	becauseitsthecup	storm	boston	bruin	juli	donat	kane	parad	meg	lebron	crawford	mayo	shaw	champion	miami	heyciara	cinco
Topic7	data	current	stanley	stanleycup	becauseitsthecup	storm	boston	bruin	juli	donat	issu	trade	angel	meg	lebron	miami	spur	appli	entertain	teacher
Topic8	data	stanley	stanleycup	becauseitsthecup	boston	parad	meg	brunch	gtgtgt	scandal										
Topic9	click	blackhawk	whackstar	presid	januari	followback	vday	valentinesday	vote	thumb	respons	valentin	obama	dir	freshcont	oscar	rockboybam	tax	girlfriend	senat
Topic10	blackhawk	whackstar	januari	vote	thumb	valentin	oscar	hawk												
Topic11	blackhawk	whackstar	januari	followback	vote	thumb	valentin	obama	grammi	strategi	eve	justin	sotu	resolut	west	shower	beyonc	option	femal	nye
Topic12	blackhawk	whackstar	presid	januari	followback	vote	thumb	valentin	obama	oscar	grammi	chichat	hawk	jan	beyonc	raven	onwardlu	superbowl	lewi	ray
Topic13	boston	angel	heat	kca	ebert	bracket	nate	snow	march	twittertalk	master	recruit	marriag	ypnchat	patrick					
Topic14	chicagonistal	appli	brunch	unlock	followback	onwardlu	snow	march	marchmad	patrick	goodfoodchi	pro	topic	transit	austin	select	gffchi	sxsw	streak	winter
Topic15	chichat	onwardlu	kca	snow	march	twittertalk	allnatur	patrick	pro	topic	transit	select	gffchi	sxsw	wayn	regret	reader	pope	pie	irish
Topic16	marathon	rain	boston	trade	appli	heat	ebert	snow	march	roger	smssummit	bomb	opportun	polic	gun	explos	draft	suspect	easter	moneymami
Topic17	rain	septemb	juli	miley	emmett	lollapalooza	grace	mixtap	que	independ	elev	bear	feat	usa	twerk	due	lolla	fourth	gunplay	
Topic18	award	juli	lollapalooza	mixtap	aug	que	bear	lolla	bieber	allscript	zimmerman	georg	djpierr	pride	race	prod	shark	phish	campaign	mtvhottest
Topic19	juli	miley	feat	prod	cyru	ilo	kaleo	ave	streetrunn	vma	nsync	taylor								
Topic20	juli	freestyl	georg																	

Table B.5: The most probable terms for topics extracted from temporal model with 10 partitions. The threshold of distribution more than 0.001 has been applied.

Topic1	emmi	smoov	smwchicago																	
Topic2	rain	content	getglu	info	syria	labor														
Topic3	djpierr	ave	pride	feat	firework	pitchfork	parad													
Topic4	zimmerman	pitchfork	campaign	wavefront	fefe	newsong	getnun	fefeondablock	freestyl	harri	kingofbop	sharknado	boppin	georg	royal					
Topic5	miami	connect	instead	constantcontact	nba	hockey	appli	trade	prom	king	heyciara	nrashow	meg	lebron	detroit					
Topic6	somosfio																			
Topic7	onwardlu	topic	heat	kca	hire	gffchi	followback	snow	pro	marchmad	select	allnatur	transit	rambler	pope	festiv	easter	sxsw	streak	patrick
Topic8	hire	snow	pope	festiv	easter	sxsw	streak	patrick	intern	goodfoodchi	irish	austin								
Topic9	rain	opportun	blackhawk	whackstar	ray	que	monday	degre	beyonc	inaug	vote	releas	inaugur	lewi	obama	tax	freshcont	goldenglob	gay	presid
Topic10	appli	onwardlu	rambler	chichat	blackhawk	whackstar	monday	degre	releas	lewi	freshcont	lanc	teachpluschicago	system	hawk	rockboybam	girlfriend	nhl	boozer	nye
Topic11	onwardlu	followback	pope	ray	beyonc	lewi	hawk	grammi	footbal	raven	justin	kelli	halftim	aye	superbowl	thumb	taylor	oscar		
Topic12	harri	followback	whackstar	beyonc	obama	freshcont	hawk	nhl	grammi	thumb	oscar	sotu	glass	ann	west	cuti	harlem	contest	data	
Topic13	chicagonistal	taam	ebert	smssummit	bomb	gno	illinoi	earth	polic	fool	websit	roger	master	louisvil	nate	suspect	internet			
Topic14	trade	ebert	smssummit	bomb	polic	roger	suspect	bostonmarathon	rsvp	version	prize	theonlinemom	prayer	explos	moneymami	flood				
Topic15	parad	stanleycup	bruin	crawford																
Topic16	parad	meg	data	stanleycup	storm	kany	letsgohawk	bruin	shaw	champion	spur	invit	appreci	court	nsa					
Topic17	campaign	que	miley	cyru	lollapalooza	vma	jennif	lolla												
Topic18	feat	que	lollapalooza	mixtap	lolla	mtvhottest	bieber	elev	kendrick	gunplay										
Topic19	happ	event																		
Topic20	costum	smatter	shutdown	program	ciw															

Table B.6: The most probable terms for topics extracted from temporal model with 20 partitions. The threshold of distribution more than 0.001 has been applied.

Topic1	lem	happ																		
Topic2	labor																			
Topic3	lollapalooza	hall	wednesday	campaign	jennif	lolla														
Topic4	justin	zimmerman	pride																	
Topic5	rain	stanley	storm	bruin	meg															
Topic6	mayo	cinco																		
Topic7	chicagonistal	trade	wednesday	boston	ebert	explos	fool	roger	master	hawk	servic									
Topic8	brand	followback	onwardlu	kca	goodfoodchi	austin	pope	streak												
Topic9	wednesday	tuesday	followback	hawk	pope	grammi	sotu	raven	whackstar	earn	act	kelli	star	aye	superbowl	thumb	illinoi	lewi	footbal	ray
Topic10	monday	holiday	whackstar	thumb	ray	cont	choic	public	hockey	releas	coupl	boozer	que	avail	mpoint	angel	goldenglob	irresist	trend	nye
Topic11	info	monday	gay	chichat	lewi	releas	que	lanc	presid	teachpluschicago	degre	project	inaug	inaugur	report	freshcont	beyonc	rockboybam	fake	manti
Topic12	googl	followback	vote	hawk	goodmorn	cuti	whackchat	valentin												
Topic13	miami	streak	heat	pro	michigan	topic	twittertalk	allnatur	transit	recruit	select	local	host	patrick						
Topic14	smssummit	flood	earth	marathon	prize	theonlinemom	cnn	moneymami												
Topic15	chicagonistal	meg	heyciara	nrashow	sale	entertain	detroit													
Topic16	wednesday	parad	spur	lebron	report	score	tip	invit	champion											
Topic17	campaign	fefe	pitchfork	royal																
Topic18	mtvhottest	que	miley	bieber	emmett	vma	feat	gunplay												
Topic19	emmi	women	smwchicago																	
Topic20	costum	ciw																		

Bibliography

[1] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[2] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.

[3] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Twitter improves seasonal influenza prediction. In *HEALTHINF*, pages 61–70, 2012.

[4] Alice E Marwick et al. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society*, 13(1):114–133, 2011.

[5] Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[6] Amr Ahmed and Eric P Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*, 2012.

[7] Loulwah AlSumait, Daniel Barbará, and Carlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 3–12. IEEE, 2008.

[8] Matthew S Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[9] Yelena Kropivnitskaya, Kristy F Tiampo, Jinhui Qin, and Michael A Bauer. The predictive relationship between earthquake intensity and tweets rate for real-time ground-motion estimation. *Seismological Research Letters*, 2017.

[10] Walid Al-Saqaf and Christian Christensen. Mainstream media power and lost orphans: The formation of twitter networks in times of conflict. 2017.

[11] Arlene E Chung, Asheley C Skinner, Stephanie E Hasty, and Eliana M Perrin. Tweeting to health: a novel mhealth intervention using fitbits and twitter to foster healthy lifestyles. *Clinical Pediatrics*, 56(1):26–32, 2017.

[12] S Hale, Devin Gaffney, and Mark Graham. Where in the world are you? geolocation and language identification in twitter. *Proceedings of ICWSM12*, pages 518–521, 2012.

[13] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5: 31, 2011.

[14] Mehran Kamkarhaghighi, Iuliia Chepurna, Somayyeh Aghababaei, and Masoud Makrehchi. Discovering credible twitter users in stock market domain. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, pages 66–72. IEEE, 2016.

[15] Pablo D Azar and Andrew W Lo. The wisdom of twitter crowds: Predicting stock market reactions to fomc meetings via twitter feeds. *The Journal of Portfolio Management*, 42(5):123–134, 2016.

[16] Johnny Torres, Gabriela Baquerizo, Carmen Vaca, and Enrique Peláez. Characterizing influential leaders of ecuador on twitter using computational intelligence. In *eDemocracy & eGovernment (ICEDEG), 2016 Third International Conference on*, pages 159–163. IEEE, 2016.

[17] B Colin Cork and Terry Eddy. The retweet as a function of electronic word-of-mouth marketing: A study of athlete endorsement activity on twitter. *International Journal of Sport Communication*, 10(1):1–16, 2017.

[18] Shiyang Gong, Juanjuan Zhang, Ping Zhao, and Xuping Jiang. Tweeting as a marketing tool–field experiment in the tv industry. *Journal of Marketing Research*, 2016.

[19] Frederik De Grove, Evelien D'heer, and Sarah Van Leuven. Where has the news gone? a network approach to secondary gatekeeping on twitter in the netherlands and belgium. In *Etmaal van de communicatiewetenschap 2016*, 2016.

[20] Axel Bruns, Brenda Moon, Felix Münch, Jan-Hinrik Schmidt, Lisa Merten, Hallvard Moe, and Sander Schwartz. News sharing on twitter: A nationally comparative study. 2016.

[21] Xialing Lin, Kenneth A Lachlan, and Patric R Spence. Exploring extreme events on social media: A comparison of user reposting/retweeting behaviors on twitter and weibo. *Computers in Human Behavior*, 65:576–581, 2016.

[22] Uuf Brajawidagda, Christopher G Reddick, and Akemi Takeoka Chatfield. Social media and urban resilience: A case study of the 2016 jakarta terror attack. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, pages 445–454. ACM, 2016.

[23] Mohammad Akbari, Xia Hu, Liqiang Nie, and Tat-Seng Chua. From tweets to wellness: Wellness event detection from twitter streams. In *AAAI*, pages 87–93, 2016.

[24] Jennifer B Unger, Patricia Escobedo, Jon-Patrick Allem, Daniel W Soto, Kar-Hai Chu, and Tess Cruz. Perceptions of secondhand e-cigarette aerosol among twitter users. *Tobacco Regulatory Science*, 2(2):146–152, 2016.

[25] Alan Steinberg, Clayton Wukich, and Hao-Che Wu. Central social media actors in disaster information networks. *International Journal of Mass Emergencies and Disasters*, 34(1):47–74, 2016.

[26] Naohiro Matsumura, Asako Miura, Masashi Komori, and Kai Hiraishi. Media mediate sentiments: Exploratory analysis of tweets posted before, during, and after the great east japan earthquake. *International Journal of Knowledge Society Research (IJKSR)*, 7(2):57–71, 2016.

[27] Kalyani Suresh and Chitra Ramakrishnan. Twittering public sentiments: A predictive analysis of pre-poll twitter popularity of prime ministerial candidates for the indian elections 2014. *Journal of Resources, Energy and Development*, 7 (2):238–254, 2016.

[28] Daniel E O'Leary. Crowd performance in prediction of the world cup 2014. *European Journal of Operational Research*, 260(2):715–724, 2017.

[29] Masoud Makrehchi. Social link recommendation by learning hidden topics. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 189–196. ACM, 2011.

[30] Kriste Krstovski, David A Smith, and Michael J Kurtz. Automatic construction of evaluation sets and evaluation of document similarity models in large scholarly retrieval systems. *arXiv preprint arXiv:1601.01611*, 2016.

[31] Jason Chuang, Christopher D Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

[32] Zhi-qiang Li, Shuai-yi Cao, and Hong-chen Guo. An improved ml-knn multi-label classification model based on feature dimensionality reduction. *DEStech Transactions on Computer Science and Engineering*, (cmee), 2016.

[33] Jessica Clark and Foster Provost. Matrix-factorization-based dimensionality reduction in the predictive modeling process: A design science perspective. 2016.

[34] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[35] Ding Zhou, Xiang Ji, Hongyuan Zha, and C Lee Giles. Topic evolution and social interactions: how authors effect research. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 248–257. ACM, 2006.

[36] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.

[37] Randall Wald, Taghi M Khoshgoftaar, Amri Napolitano, and Chris Sumner. Using twitter content to predict psychopathy. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 394–401. IEEE, 2012.

[38] Daisuke Yokomoto, Kensaku Makita, Hiroko Suzuki, Daichi Koike, Takehito Utsuro, Yasuhide Kawada, and Tomohiro Fukuhara. Lda-based topic modeling in labeling blog posts with wikipedia entries. In *Web technologies and applications*, pages 114–124. Springer, 2012.

[39] D Teja Santosh, K Sudheer Babu, SDV Prasad, and A Vivekananda. Opinion mining of online product reviews from traditional lda topic clusters using feature ontology tree and sentiwordnet. 2016.

[40] Xin Guo, Yang Xiang, Qian Chen, Zhenhua Huang, and Yongtao Hao. Lda-based online topic detection using tensor factorization. *Journal of Information Science*, page 0165551512473066, 2013.

[41] Jey Han Lau, Nigel Collier, and Timothy Baldwin. On-line trend analysis with topic models:\# twitter trends detection topic model online. In *COLING*, pages 1519–1534. Citeseer, 2012.

[42] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM, 2010.

[43] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

[44] Yu Wang, Eugene Agichtein, and Michele Benzi. Tm-lda: efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–131. ACM, 2012.

[45] Abram Hindle, Michael W Godfrey, and Richard C Holt. What's hot and what's not: Windowed developer topic analysis. In *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*, pages 339–348. IEEE, 2009.

[46] Erich Schubert, Michael Weiler, and Hans-Peter Kriegel. Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds.

In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 871–880. ACM, 2014.

[47] Carmen K Vaca, Amin Mantrach, Alejandro Jaimes, and Marco Saerens. A time-based collective factorization for topic discovery and monitoring in news. In *Proceedings of the 23rd international conference on World wide web*, pages 527–538. International World Wide Web Conferences Steering Committee, 2014.

[48] Chung-Hong Lee, Tzan-Feng Chien, and Hsin-Chang Yang. An automatic topic ranking approach for event detection on microblogging messages. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 1358–1363. IEEE, 2011.

[49] Hanna M Wallach, David M Mimno, and Andrew McCallum. Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981, 2009.

[50] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic significance ranking of lda generative models. In *Machine Learning and Knowledge Discovery in Databases*, pages 67–82. Springer, 2009.

[51] Saptarshi Ghosh, Muhammad Bilal Zafar, Parantapa Bhattacharya, Naveen Sharma, Niloy Ganguly, and Krishna Gummadi. On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1739–1744. ACM, 2013.

[52] Shing Doong and Daniel Chung. Exploring time series spectral features in viral hashtags prediction. In *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[53] George Veletsianos. Toward a generalizable understanding of twitter and social media use across moocs: who participates on mooc hashtags and in what ways? *Journal of Computing in Higher Education*, pages 1–16, 2017.

[54] Kenton White, Guichong Li, and Nathalie Japkowicz. Sampling online social networks using coupling from the past. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 266–272. IEEE, 2012.

[55] Gi Woong Yun, Morin David, Sanghee Park, Claire Youngnyo Joa, Brett Labbe, Jongsoo Lim, Sooyoung Lee, and Daewon Hyun. Social media and flu: Media twitter accounts as agenda setters. *International journal of medical informatics*, 91:67–73, 2016.

[56] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2006.

[57] Somayyeh Aghababaei and Masoud Makrehchi. Temporal topic inference for trend prediction. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 877–884. IEEE, 2015.

[58] Emily M Cody, Andrew J Reagan, Peter Sheridan Dodds, and Christopher M Danforth. Public opinion polling with twitter. *arXiv preprint arXiv:1608.02024*, 2016.

[59] Iuliia Chepurna, Somayyeh Aghababaei, and Masoud Makrehchi. How to predict social trends by mining user sentiments. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 270–275. Springer, 2015.

[60] Shawndra Hill and Noah Ready-Campbell. Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, 15(3):73–102, 2011.

[61] Gang Wang, Tianyi Wang, Bolun Wang, Divya Sambasivan, Zengbin Zhang, Haitao Zheng, and Ben Y Zhao. Crowds on wall street: Extracting value from collaborative investing platforms. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 17–30. ACM, 2015.

[62] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*, 2013.

[63] Muhammad Bilal Zafar, Parantapa Bhattacharya, Niloy Ganguly, Krishna P Gummadi, and Saptarshi Ghosh. Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream. *ACM Transactions on the Web (TWEB)*, 9(3):12, 2015.

[64] Manish Gaurav, Amit Srivastava, Anoop Kumar, and Scott Miller. Leveraging candidate popularity on twitter to predict election outcome. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*, page 7. ACM, 2013.

[65] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. On the bias of bfs (breadth first search). In *Teletraffic Congress (ITC), 2010 22nd International*, pages 1–8. IEEE, 2010.

[66] Danni Wang, Peter Hom, Rodger Griffeth, and Jeffrey K Sager. finothing endures but changefi: Investigating temporal dynamics within a turnover model. In *Academy of Management Proceedings*, volume 2015, page 15237. Academy of Management, 2015.

[67] Satyabrata Aich, Hee-Cheol Kim, Mangal Sain, and Bijay Bhaskar Deo. Analyzing stock price changes using event related twitter feeds. In *Advanced Communication Technology (ICACT), 2017 19th International Conference on*, pages 484–487. IEEE, 2017.

[68] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–613. ACM, 2013.

[69] Linna Li, Michael F Goodchild, and Bo Xu. Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *Cartography and Geographic Information Science*, 40(2):61–77, 2013.

[70] Liangjie Hong, Aziz S Doumith, and Brian D Davison. Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 557–566. ACM, 2013.

[71] Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. Inferring the socioeconomic status of social media users based on behaviour and language. In *European Conference on Information Retrieval*, pages 689–695. Springer, 2016.

[72] David Abrahamsen. The psychology of crime. 1960.

[73] S Kirson Weinberg. Theories of criminality and problems of prediction. *The Journal of Criminal Law, Criminology, and Police Science*, pages 412–424, 1954.

[74] John Eck, Spencer Chainey, James Cameron, and R Wilson. Mapping crime: Understanding hotspots. 2005.

[75] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

[76] Xiaofeng Wang and Donald E Brown. The spatio-temporal modeling for criminal incidents. *Security Informatics*, 1(1):1–17, 2012.

[77] Yifei Xue and Donald E Brown. Spatial analysis with preference specification of latent decision makers for criminal event prediction. *Decision support systems*, 41(3):560–573, 2006.

[78] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493), 2011.

[79] Adam Boessen George E. Tita. 9 social networks and the ecology of crime: Using social network data to understand the spatial distribution of crime. pages 128–143, 2012.

[80] John R Hipp, Carter T Butts, Ryan Acton, Nicholas N Nagle, and Adam Boessen. Extrapolative simulation of neighborhood networks based on population spatial distribution: Do they predict crime? *Social Networks*, 35(4):614–625, 2013.

[81] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 427–434. ACM, 2014.

[82] Martin Traunmueller, Giovanni Quattrone, and Licia Capra. Mining mobile phone data to investigate urban crime theories at scale. In *Social Informatics*, pages 396–411. Springer, 2014.

[83] David Weisburd and Lorraine Green. Defining the street-level drug market. 1994.

[84] David Weisburd, Anthony A Braga, Elizabeth R Groff, and Alese Wooditch. Can hot spots policing reduce crime in urban areas? an agent-based simulation. *Criminology*, 55(1):137–173, 2017.

[85] Lyria Bennett Moses and Janet Chan. Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Policing and Society*, pages 1–17, 2016.

[86] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80. IEEE, 2012.

[87] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[88] Nick Malleson and Martin A Andresen. The impact of using social media data in crime rate calculations: shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42(2):112–121, 2015.

[89] Xinyu Chen, Youngwoon Cho, and Suk Young Jang. Crime prediction using twitter sentiment and weather. In *Systems and Information Engineering Design Symposium (SIEDS), 2015*, pages 63–68. IEEE, 2015.

[90] Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. Automatic targeted-domain spatiotemporal event detection in twitter. *GeoInformatica*, 20(4):765–795, 2016.

[91] Duc T Nguyen and Jai E Jung. Real-time event detection for online behavioral analysis of big social data. *Future Generation Computer Systems*, 66:137–145, 2017.

[92] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[93] Teruhiko Teraoka. Organization and exploration of heterogeneous personal data collected in daily life. *Human-Centric Computing and Information Sciences*, 2(1): 1, 2012.

[94] Atsushi Sato, Runhe Huang, and Neil Y Yen. Design of fusion technique-based mining engine for smart business. *Human-centric Computing and Information Sciences*, 5(1):1, 2015.

[95] Matko Bošnjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. Twitterecho: a distributed focused crawler to support open research with twitter data. pages 1233–1240, 2012.

[96] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xiaoming Li. Topical keyphrase extraction from twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 379–388. Association for Computational Linguistics, 2011.

[97] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 575–590. ACM, 2012.

[98] Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. Two 1% s donfit make a whole: Comparing simultaneous samples from twitterfis streaming api. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 75–83. Springer, 2014.

[99] Swit Phuvipadawat and Tsuyoshi Murata. Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123. IEEE, 2010.

[100] Mustafa Sofean and Matthew Smith. A real-time architecture for detection of diseases using social networks: design, implementation and evaluation. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 309–310. ACM, 2012.

[101] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[102] Vasileios Lampos, Daniel Preotiuc-Pietro, and Trevor Cohn. A user-centric model of voting intention from social media. In *ACL (1)*, pages 993–1003, 2013.

[103] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[104] Eric Zivot and Jiahui Wang. Rolling analysis of time series. In *Modeling Financial Time Series with S-Plus®*, pages 299–346. Springer, 2003.

[105] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[106] Sandeep Mishra. Crime drop of the 1990s. *The Encyclopedia of Criminology and Criminal Justice*, 2014.

[107] Craig A Anderson. Temperature and aggression: effects on quarterly, yearly, and city rates of violent and nonviolent crime. *Journal of personality and social psychology*, 52(6):1161, 1987.

[108] Masoud Makrehchi. *Feature ranking for text classifiers*. PhD thesis, University of Waterloo, 2007.

[109] Steven Raphael and Rudolf Winter-Ebmer. Identifying the effect of unemployment on crime*. *Journal of Law and Economics*, 44(1):259–283, 2001.

[110] Lawrence E Cohen and Marcus Felson. Social change and crime rate trends: A routine activity approach. *American sociological review*, pages 588–608, 1979.

[111] Cıcero dos Santos, Victor Guimaraes, RJ Niterói, and Rio de Janeiro. Boosting named entity recognition with neural character embeddings. In *Proceedings of NEWS 2015 The Fifth Named Entities Workshop*, page 25, 2015.

[112] Duyu Tang, Bing Qin, and Ting Liu. Deep learning for sentiment analysis: successful approaches and future challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(6):292–303, 2015.

[113] Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. Deep learning for chinese word segmentation and pos tagging. In *EMNLP*, pages 647–657, 2013.